PAT 498/598 (Winter 2025)

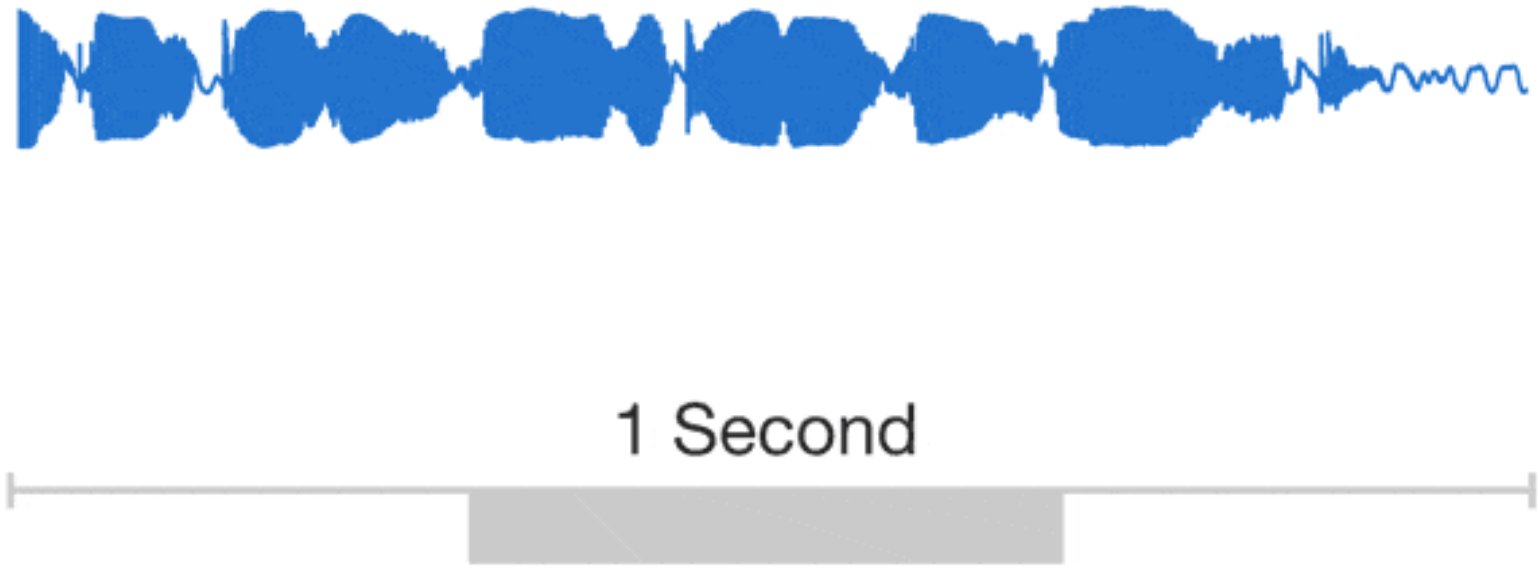
Music & AI

Lecture 17: Latent-based Music Generation

Instructor: Hao-Wen Dong

SCHOOL OF MUSIC, THEATRE & DANCE
PERFORMING ARTS TECHNOLOGY
UNIVERSITY OF MICHIGAN

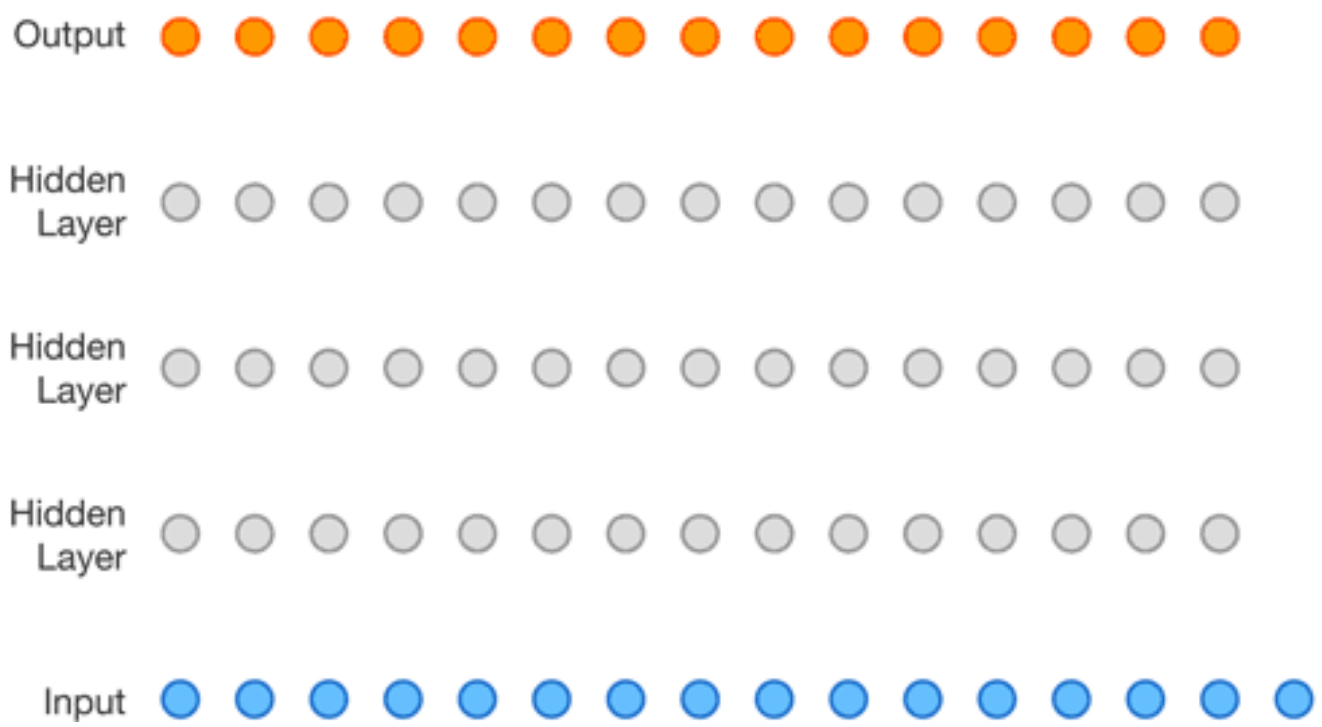
(Recap) Generating Waveforms using a Neural Network



(Source: van den Oord et al., 2016)

Aaron van den Oord and Sander Dieleman, "WaveNet: A generative model for raw audio," *DeepMind Blog*, September 8, 2016.

(Recap) WaveNet (van den Oord et al., 2016)



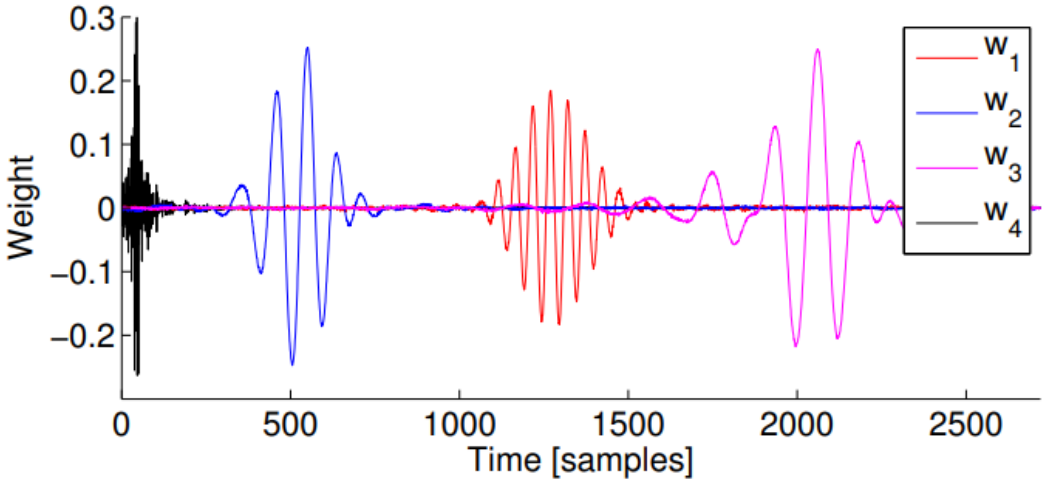
(Source: van den Oord et al., 2016)

A convolutional neural network for raw waveform generation

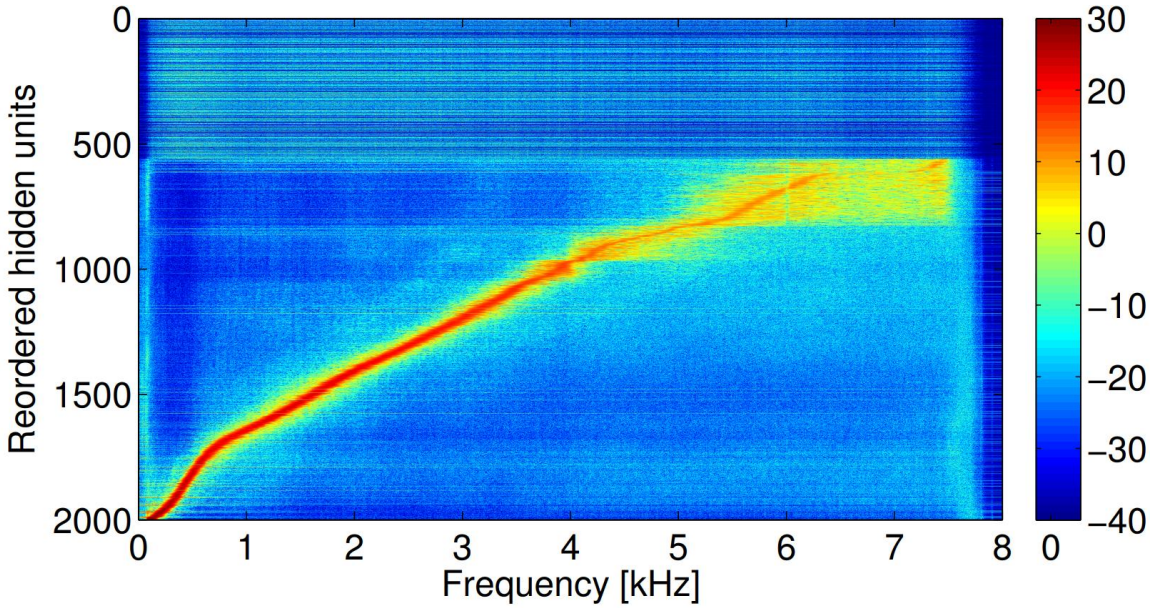
Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *ICML*, 2016.

(Recap) 1D CNNs & Fourier Transform

Convolution kernels learned

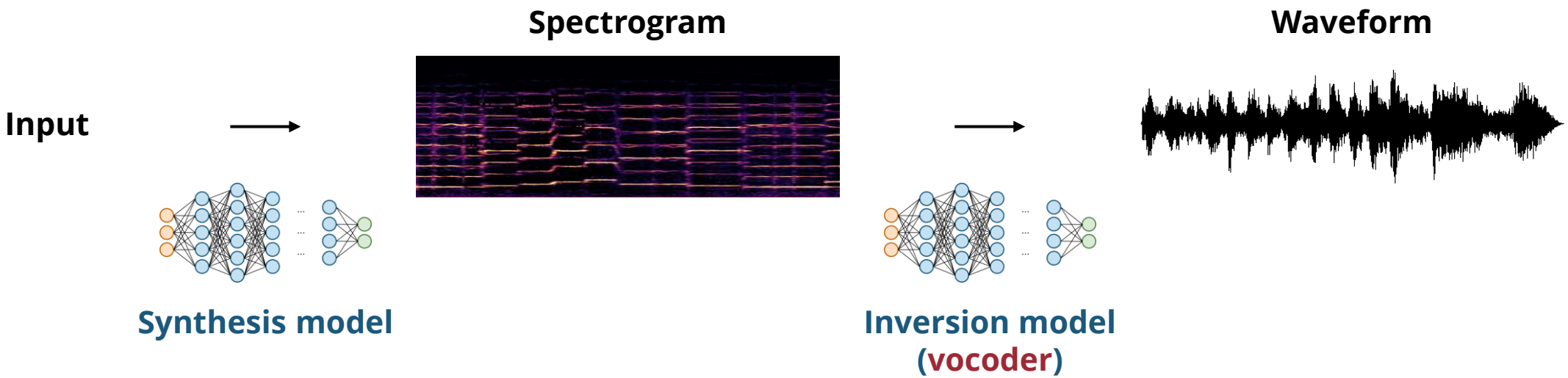


Peak frequency detected by the learned kernels

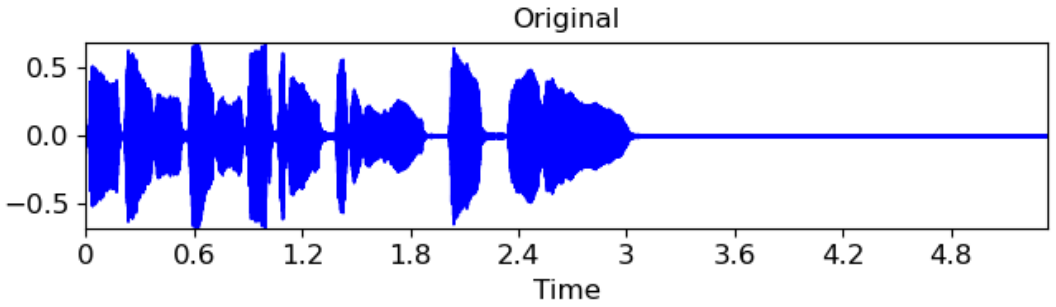


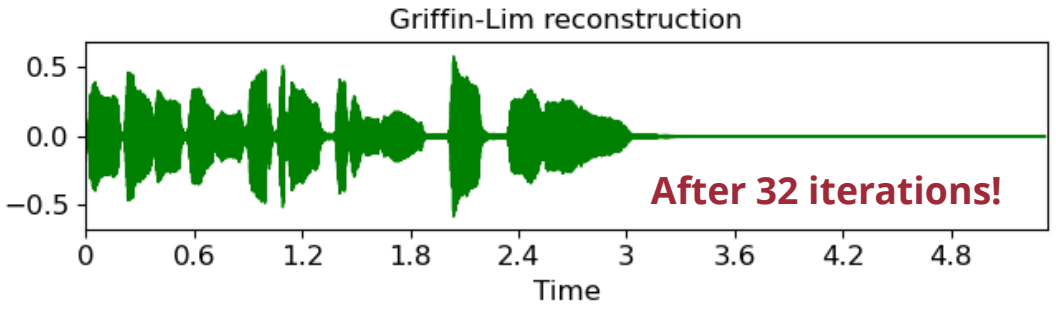
Pavel Golik, Zoltán Tüske, Ralf Schlüter, and Hermann Ney, "Convolutional Neural Networks for Acoustic Modeling of Raw Time Signal in LVCSR," *INTERSPEECH*, 2015.

(Recap) Frequency-domain Audio Synthesis



(Recap) Griffin-Lim Algorithm (Griffin & Lim, 1984)





(Source: librosa documentation)

Given a magnitude-only STFT matrix

↓

Randomly initialize the phase

↓

$$y' = \arg\min_{y} (M - \mathrm{STFT}(y))^2$$

Find the signal y that minimize the MSE between the input and STFT(y)

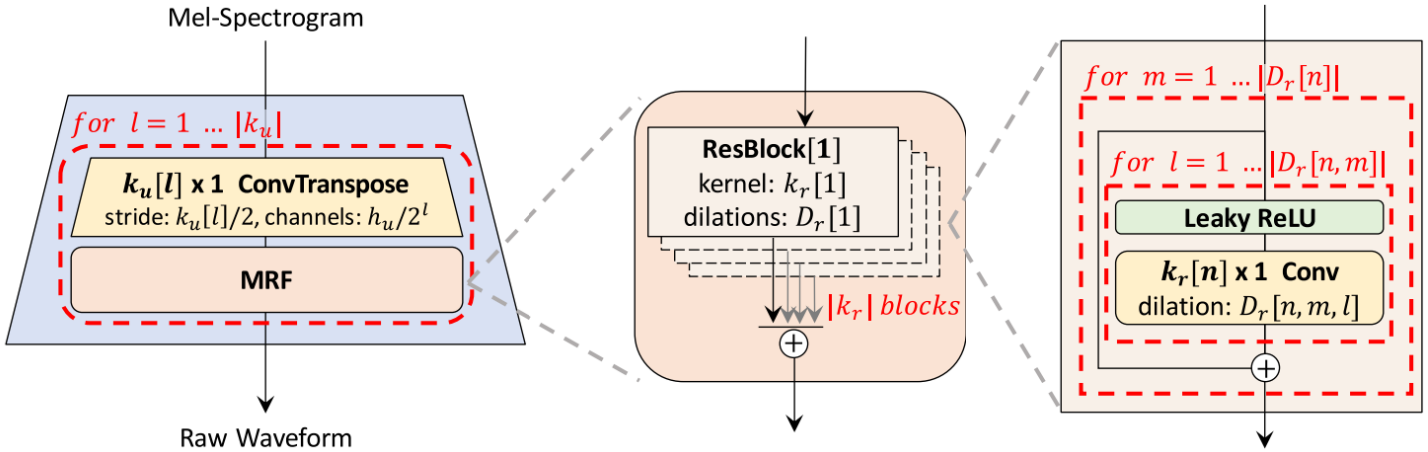
↓

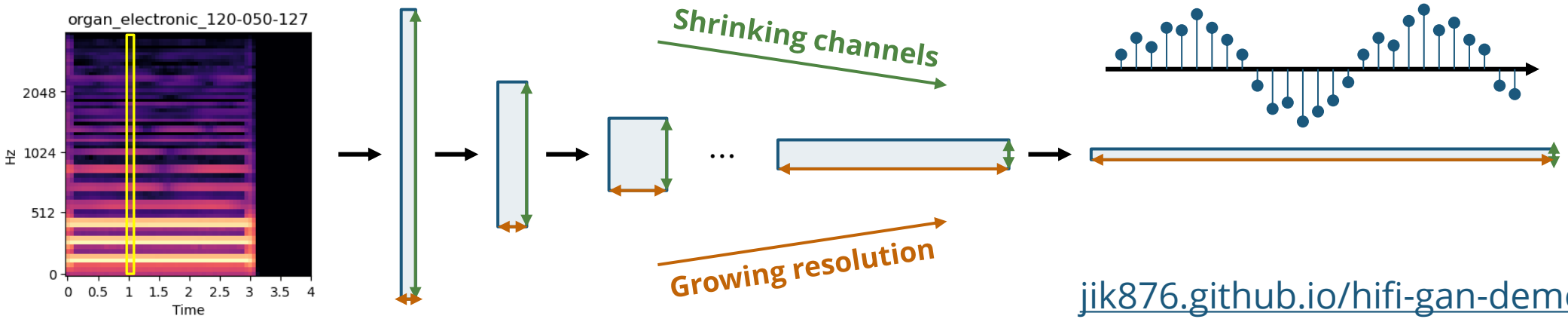
$$M' = \mathrm{STFT}(y')$$

Find the STFT of the signal y

Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE TASSP*, 1984.
librosa.org/doc/main/generated/librosa.griffinlim.html

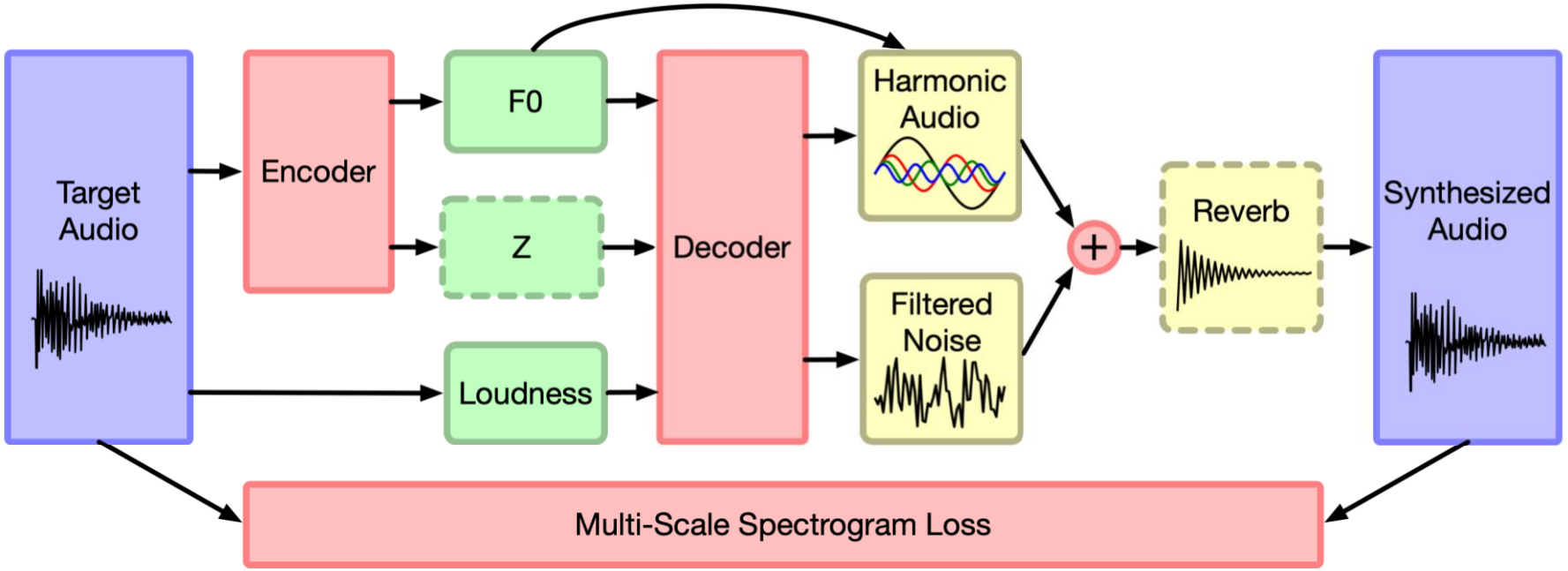
(Recap) Hifi-GAN (Kong et al., 2020)





jik876.github.io/hifi-gan-demo

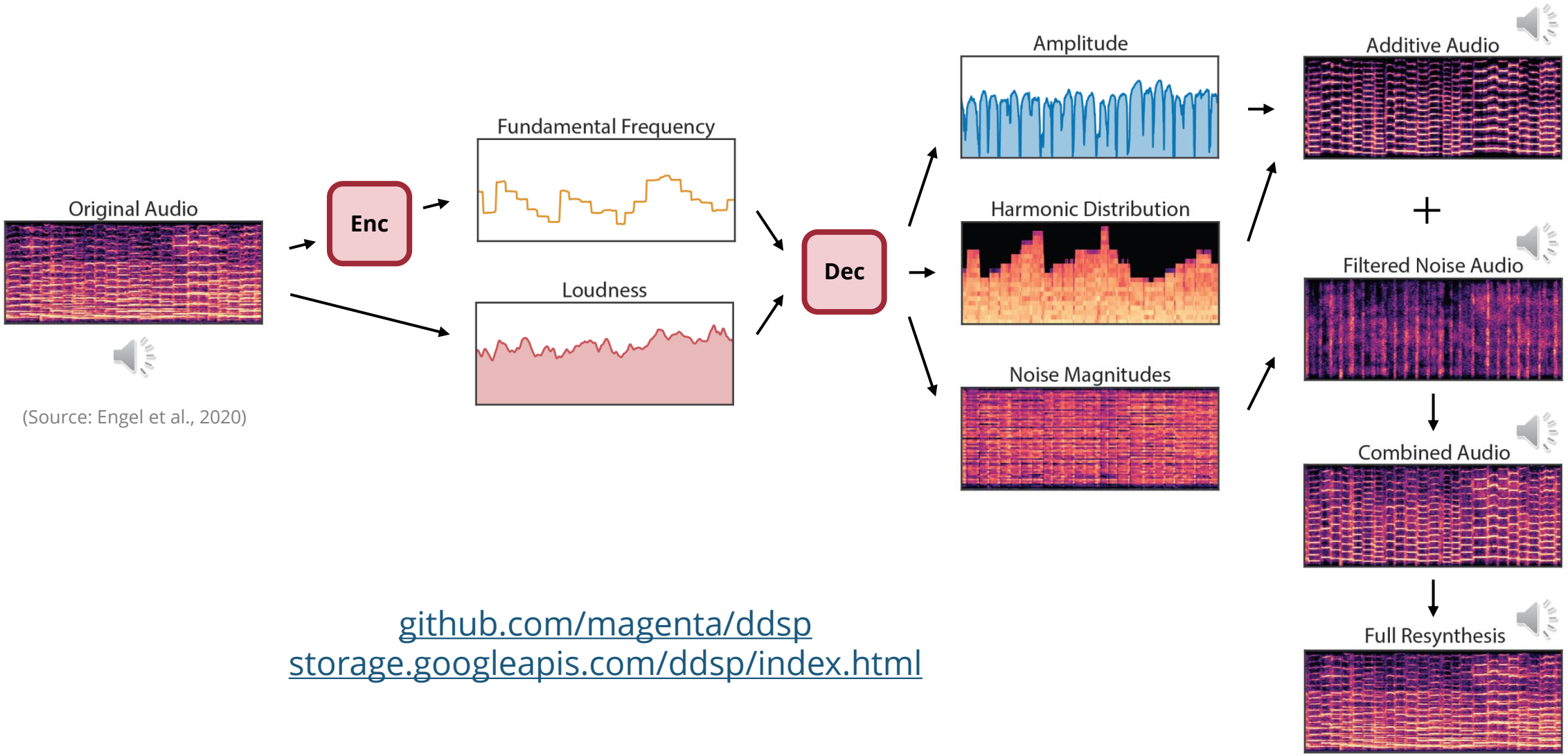
(Recap) Differentiable DSP (DDSP) (Engel et al., 2020)



(Source: Engel et al., 2020)

Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, Adam Roberts, "DDSP: Differentiable Digital Signal Processing," *ICLR*, 2020.

(Recap) Differentiable DSP (DDSP) (Engel et al., 2020)



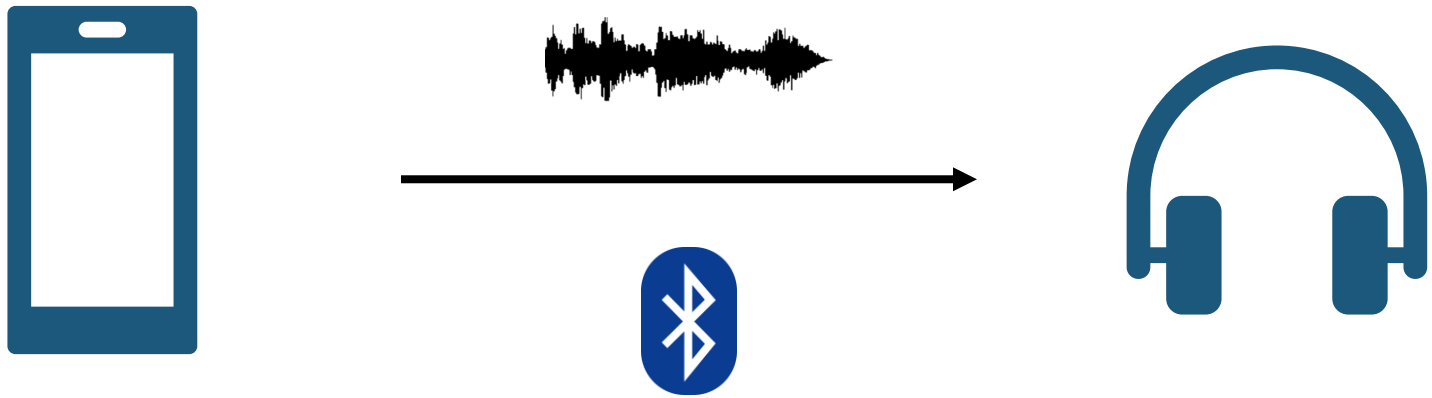
(Source: Engel et al., 2020)

github.com/magenta/ddsp
storage.googleapis.com/ddsp/index.html

Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, Adam Roberts, "DDSP: Differentiable Digital Signal Processing," *ICLR*, 2020.

Neural Codecs

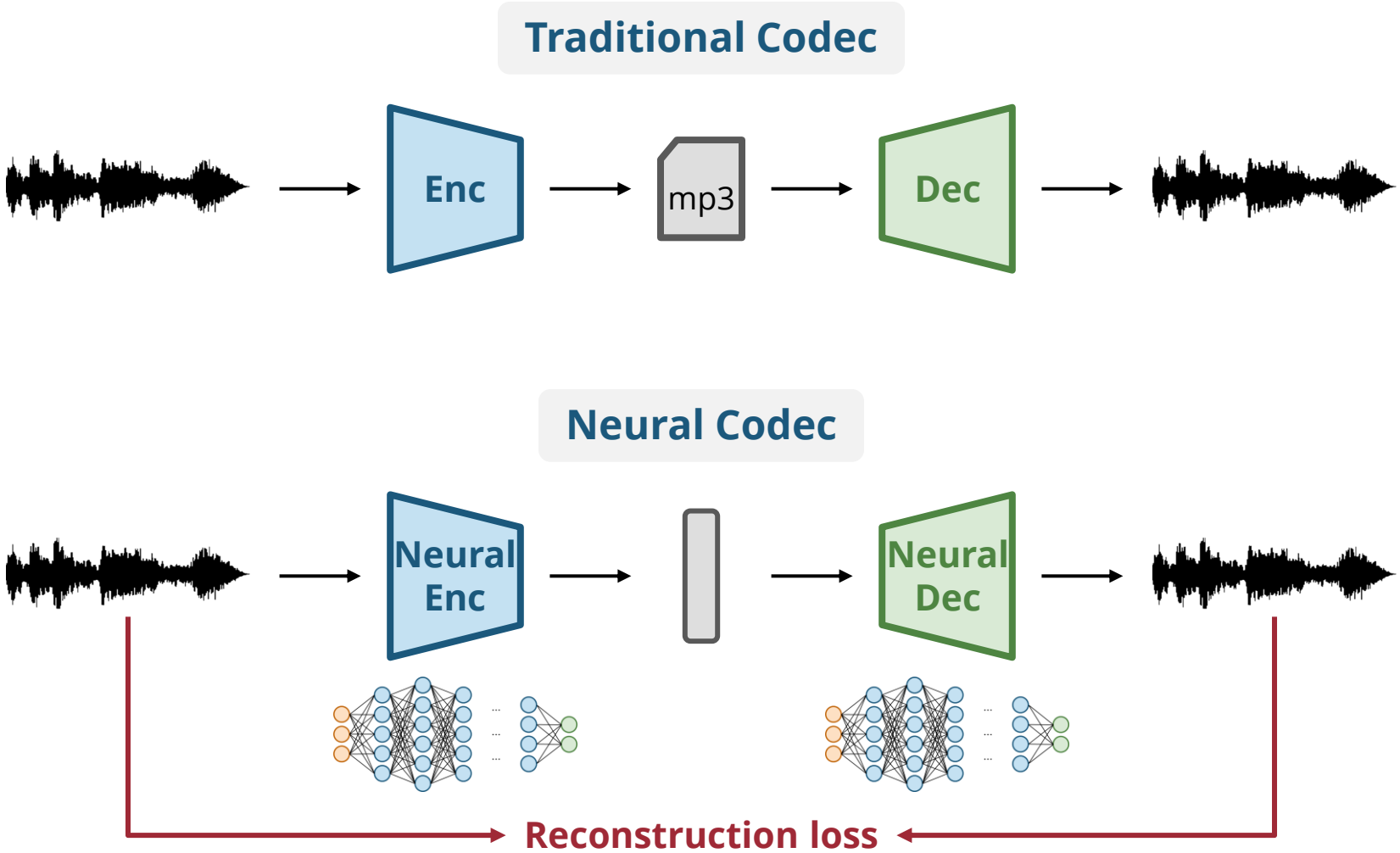
What is a Codec?



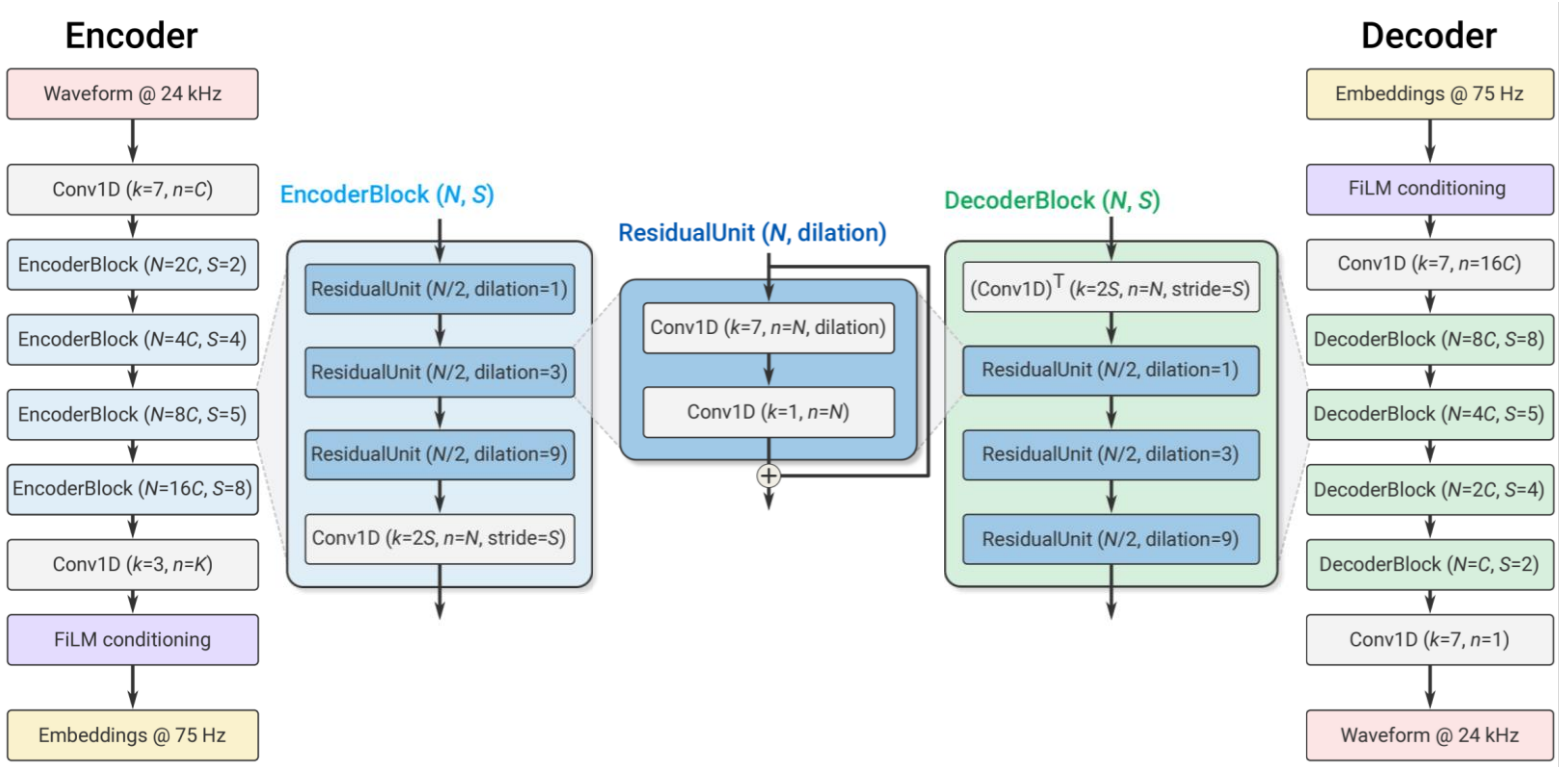
SBC, AAC, aptX,
aptX HD, LDAC

What is a Codec?

Neural Codec



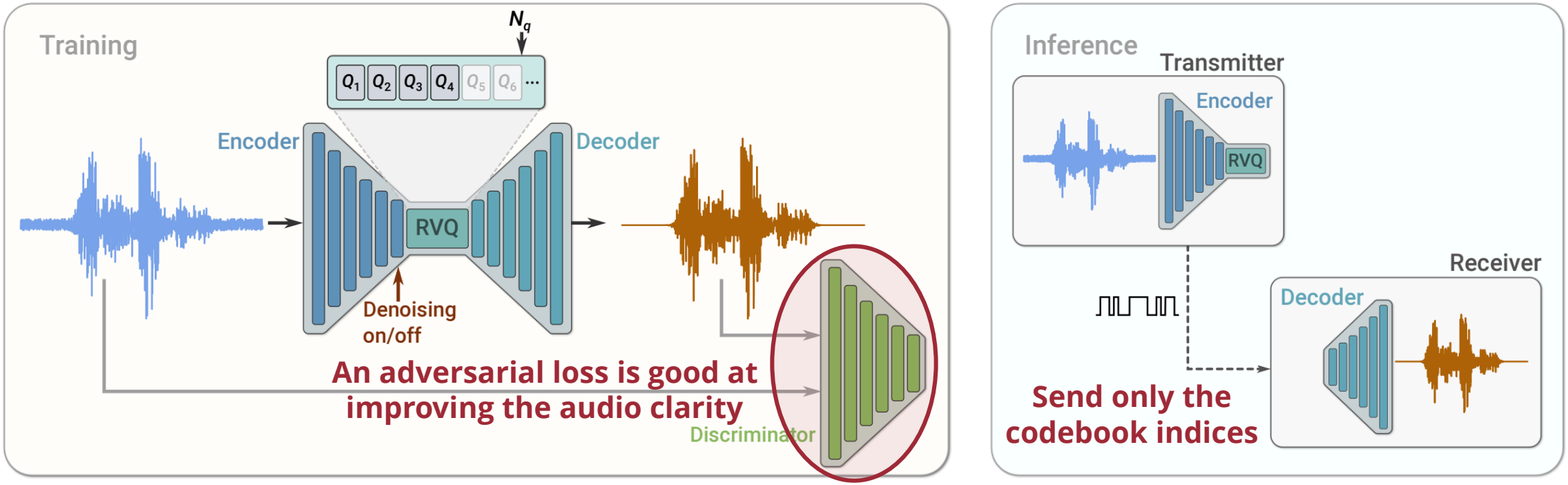
SoundStream (Zeghidour et al., 2021)



(Source: Zeghidour et al., 2021)

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "SoundStream: An End-to-End Neural Audio Codec," *TASLP*, 2021.

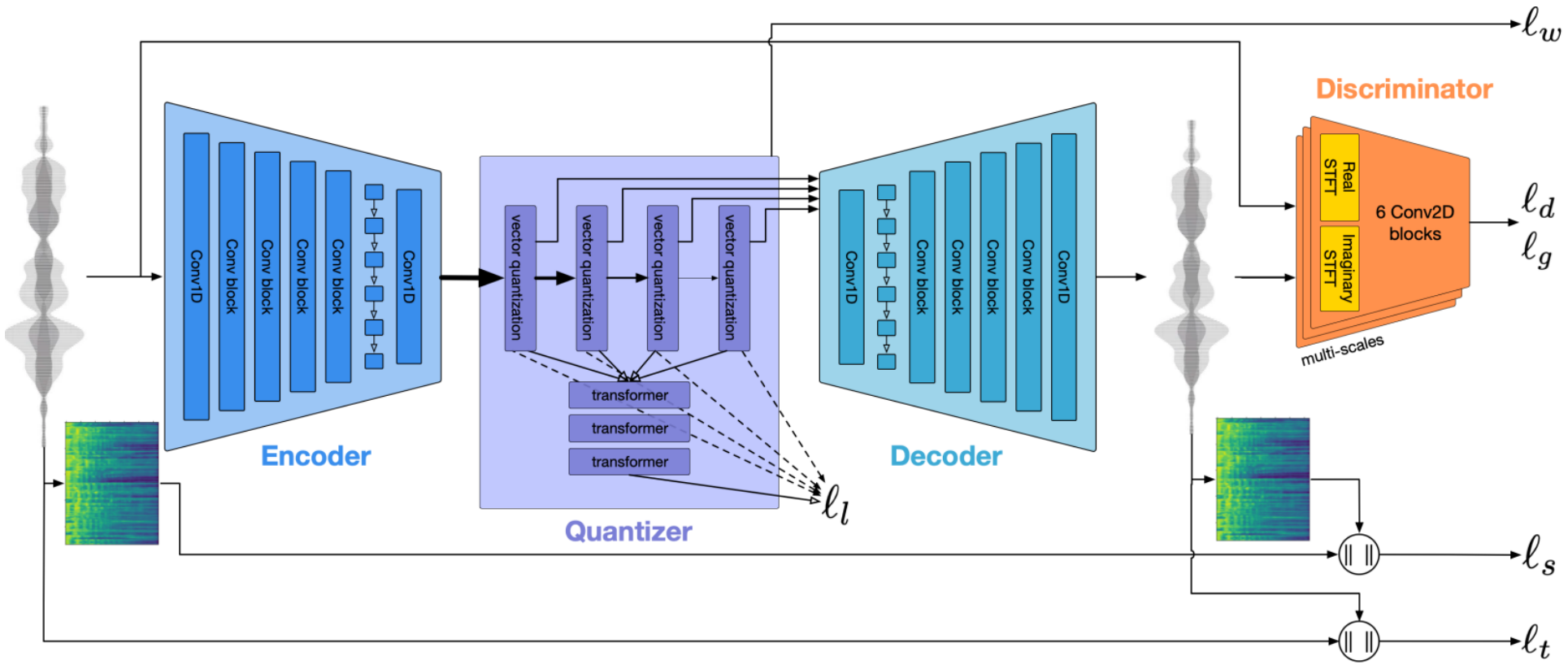
SoundStream (Zeghidour et al., 2021)



(Source: Zeghidour et al., 2021)

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "SoundStream: An End-to-End Neural Audio Codec," *TASLP*, 2021.

EnCodec (Défossez et al., 2022)



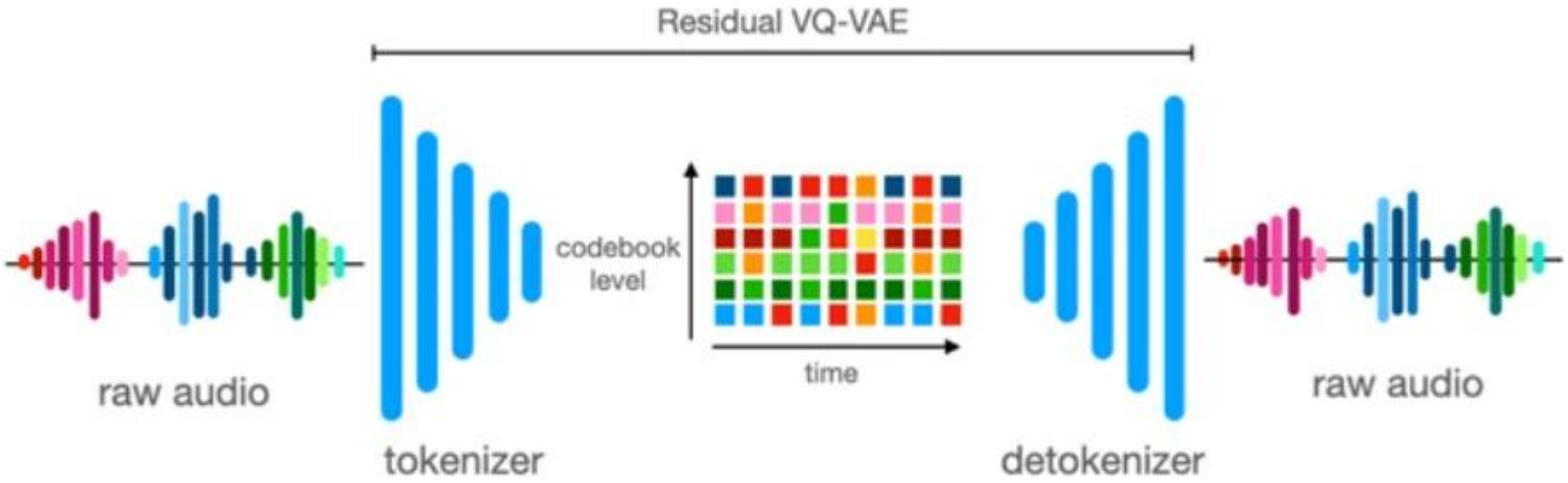
(Source: Défossez et al., 2022)

ai.honu.io/papers/encodec/samples.html

github.com/facebookresearch/encodec

Alexandre Défossez, Jade Copet, Gabriel Synnaeve and Yossi Adi, "High Fidelity Neural Audio Compression," *arXiv preprint arXiv:2210.13438*, 2022.

Descript Audio Codec (Kumar et al., 2023)



(Source: Kumar et al., 2023)

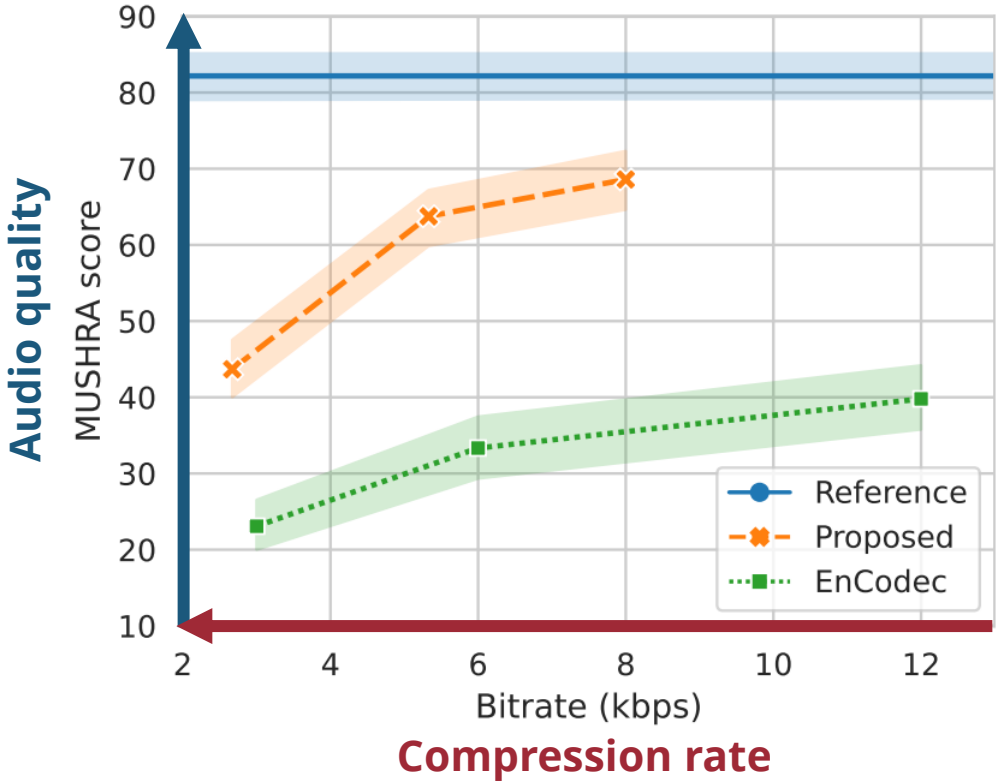
descript.notion.site/Descript-Audio-Codec-11389fce0ce2419891d6591a68f814d5

github.com/descriptinc/descript-audio-codec

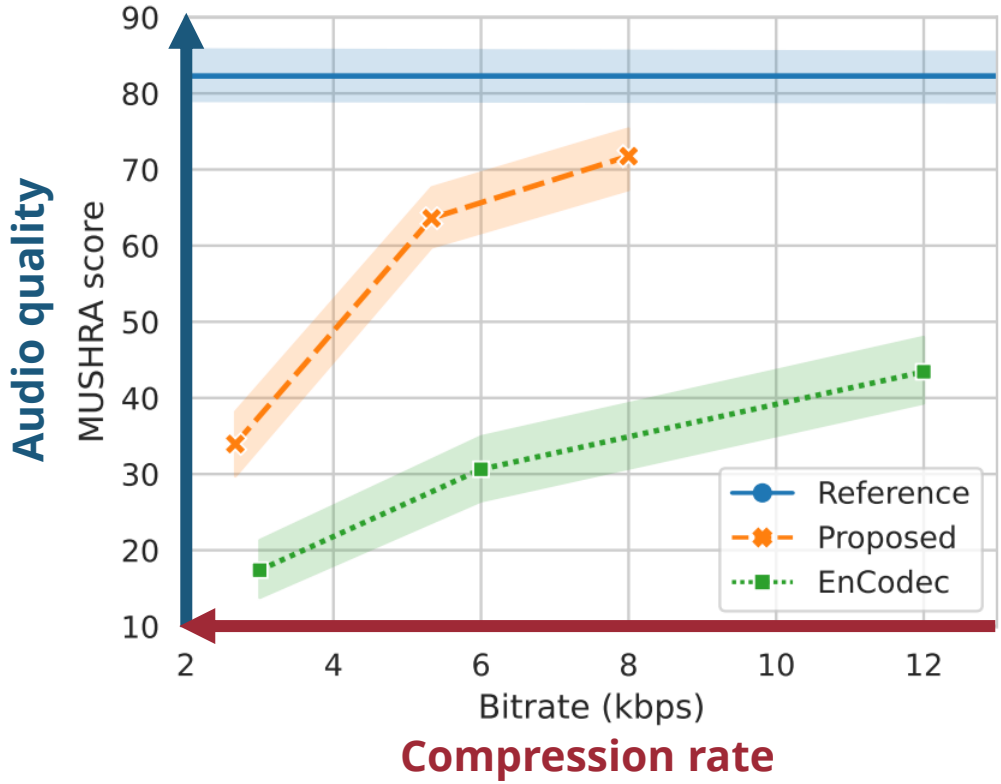
Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-Fidelity Audio Compression with Improved RVQGAN," *NeurIPS*, 2023.

Descript Audio Codec (Kumar et al., 2023)

Listening Test Results @ 44.1 kHz



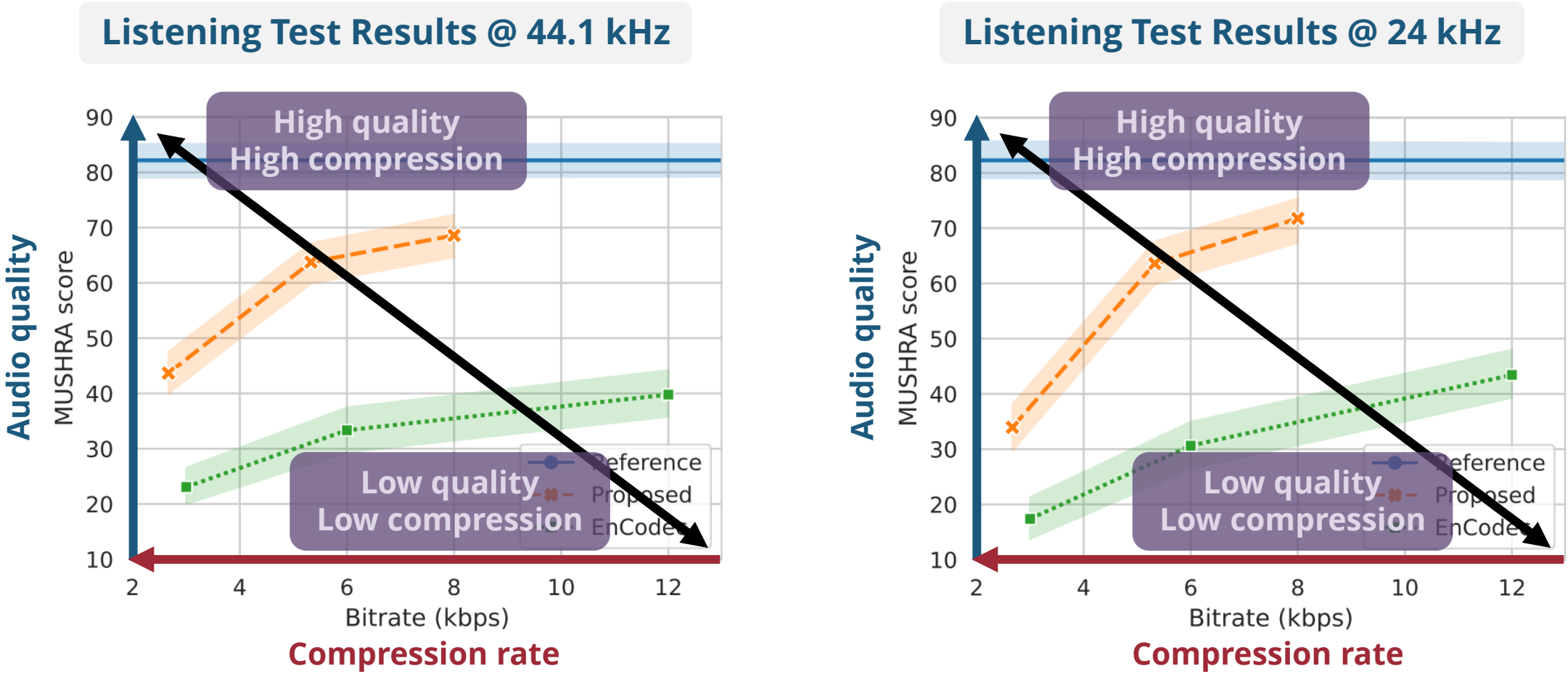
Listening Test Results @ 24 kHz



(Source: Kumar et al., 2023)

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-Fidelity Audio Compression with Improved RVQGAN," *NeurIPS*, 2023.

Descript Audio Codec (Kumar et al., 2023)



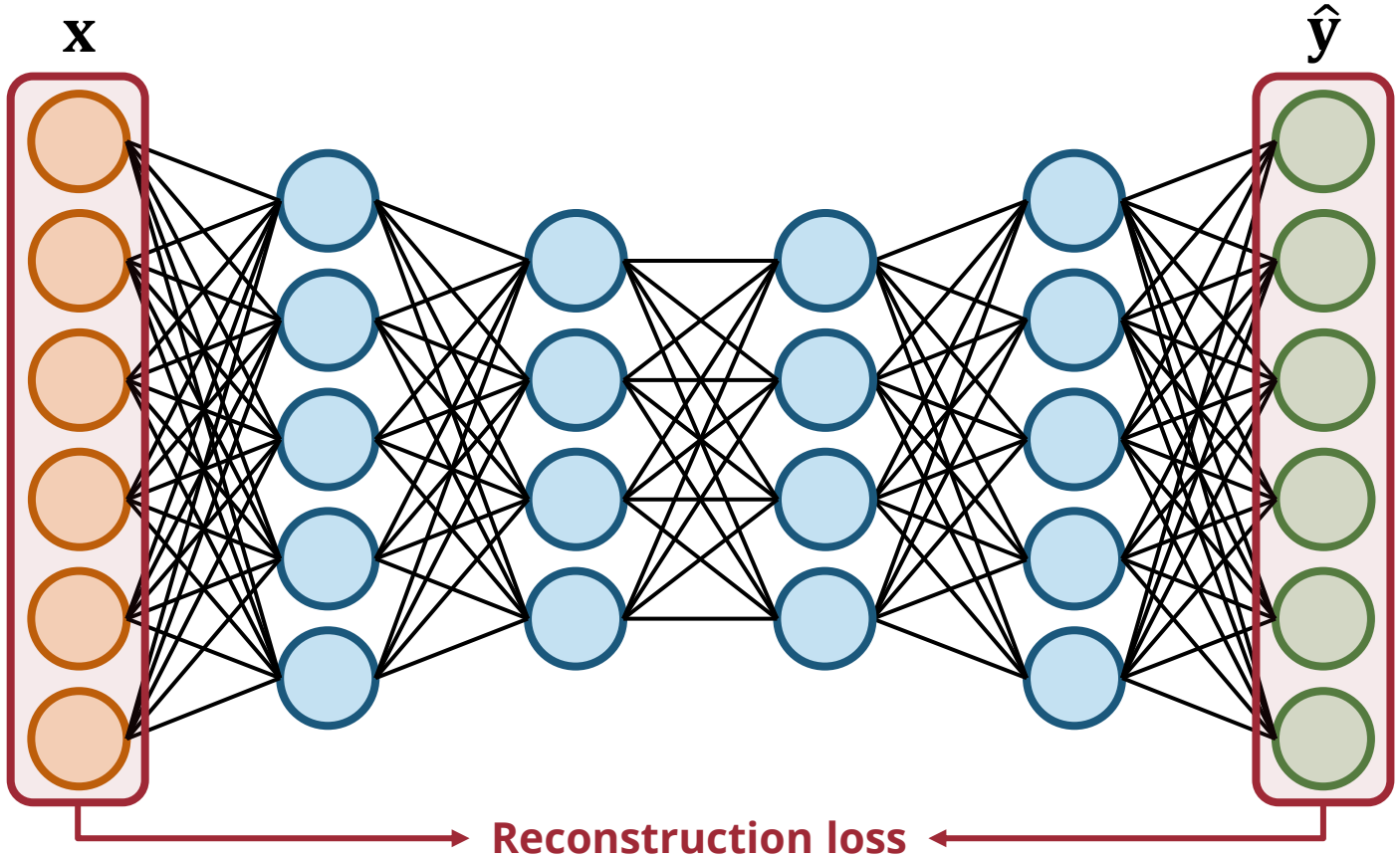
(Source: Kumar et al., 2023)

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-Fidelity Audio Compression with Improved RVQGAN," *NeurIPS*, 2023.

Latent Diffusion Models

(Recap) Autoencoders

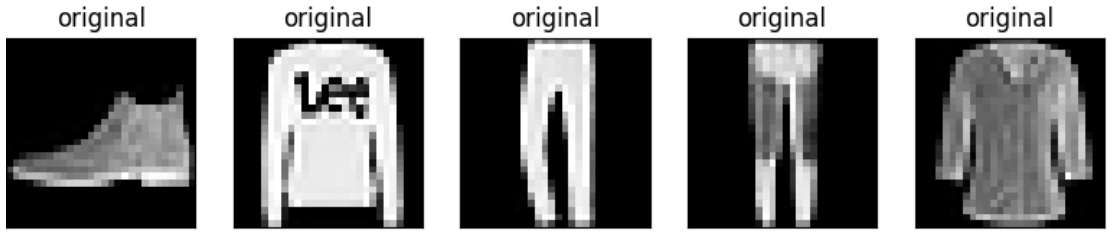
- A neural network where the **input and output are the same**



(Recap) Autoencoders – Reconstruction Examples

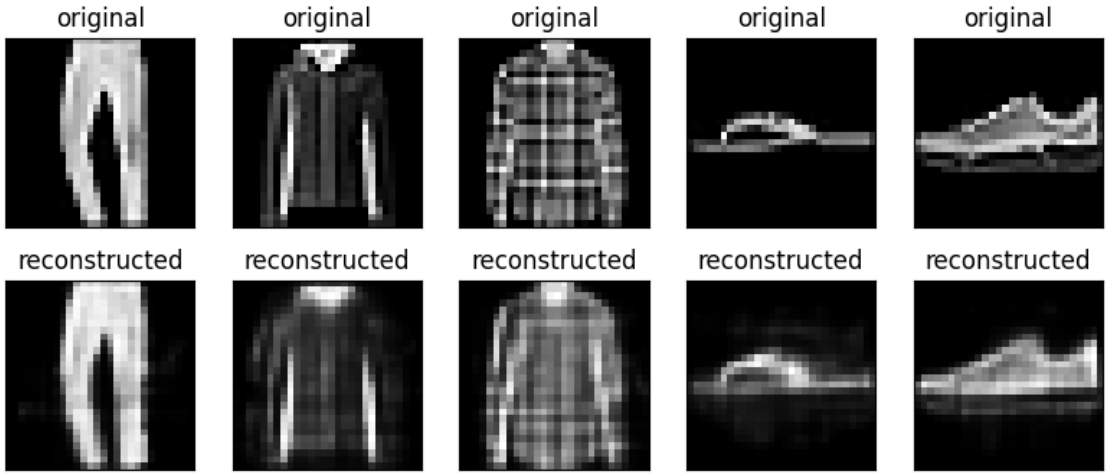
Original

Reconstructed



Original

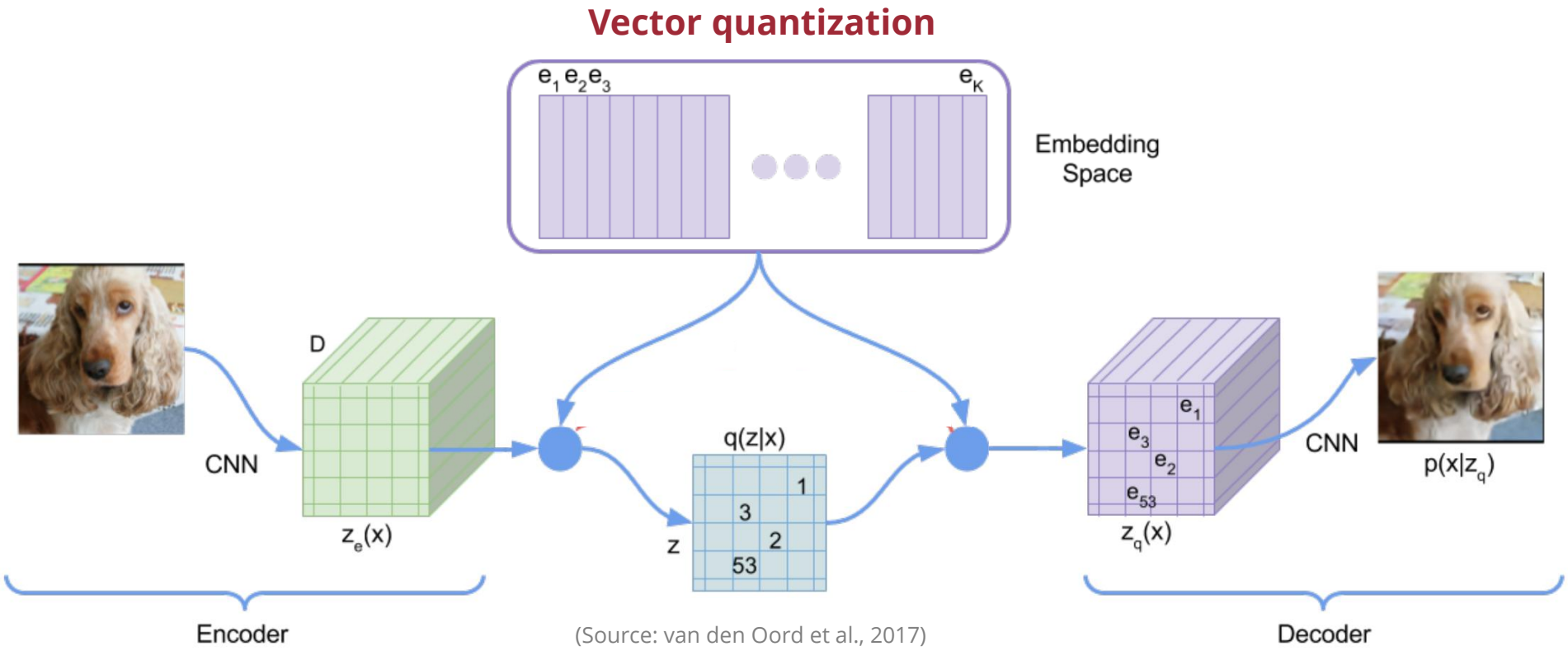
Reconstructed



(Source: tensorflow.org)

tensorflow.org/tutorials/generative/autoencoder

Vector-Quantized VAEs (VQVAEs)

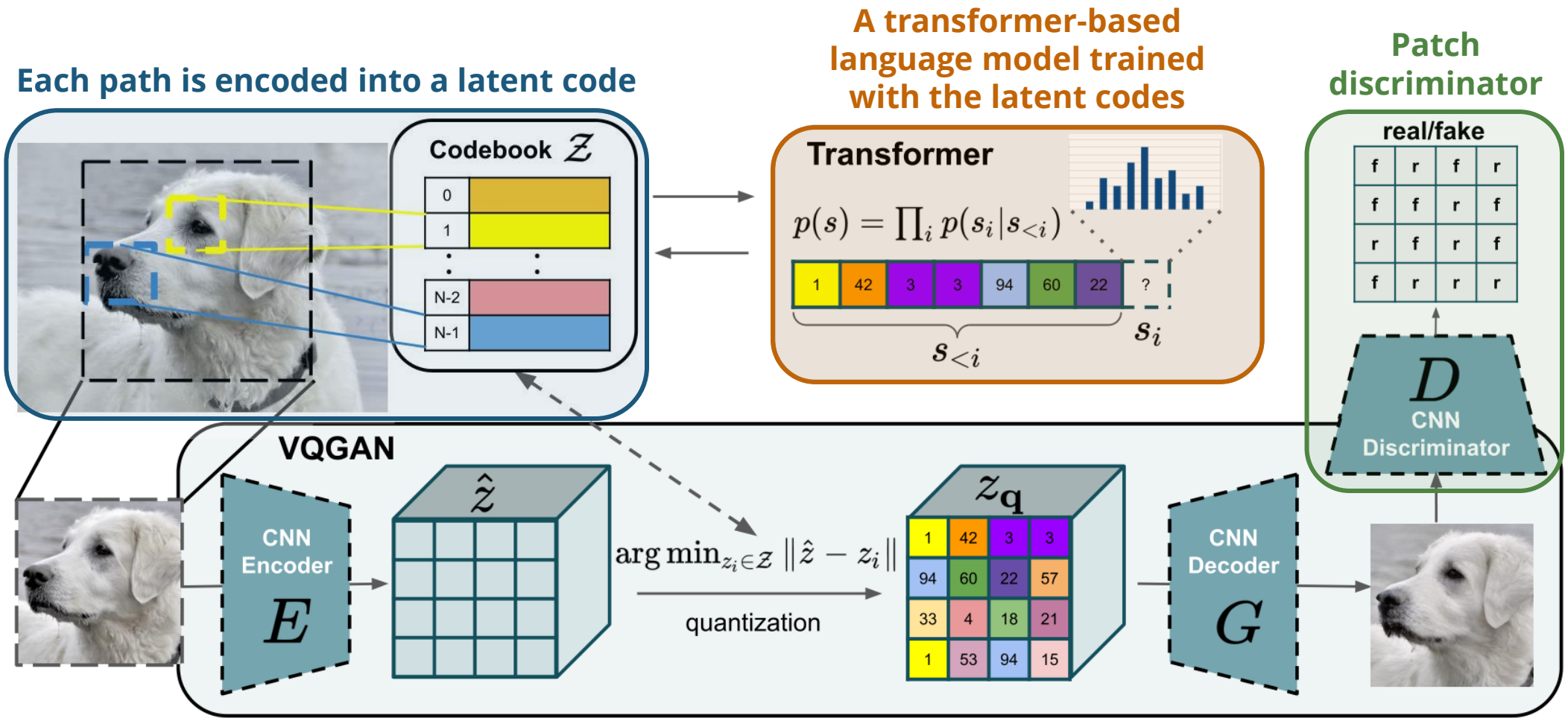


(Source: van den Oord et al., 2017)

Allow only a fixed number of vectors to be used in the bottleneck layer

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural Discrete Representation Learning," *NeurIPS*, 2017.

VQGAN



(Source: Esser et al., 2021)

A VQGAN is a VQVAE equipped with adversarial loss

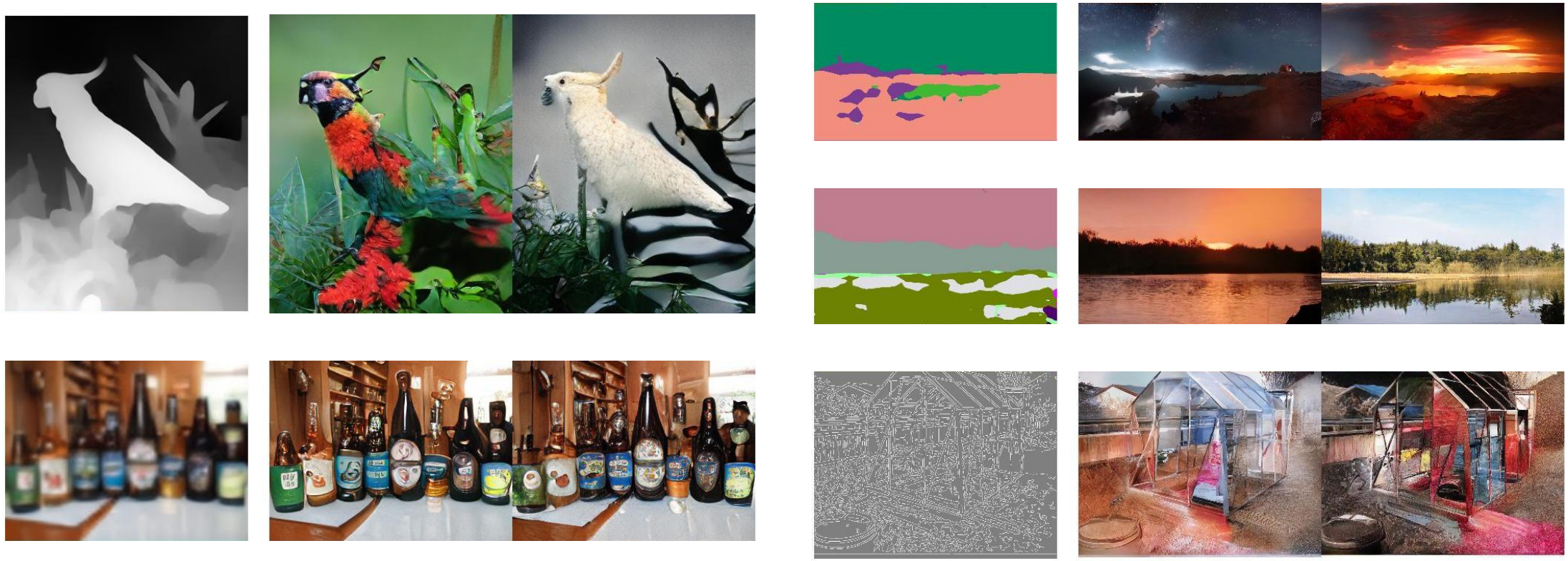
Patrick Esser, Robin Rombach, and Björn Ommer, "Taming Transformers for High-Resolution Image Synthesis," *CVPR*, 2021.

VQGAN: Conditional Generation

(Source: Esser et al., 2021)

Patrick Esser, Robin Rombach, and Björn Ommer, "Taming Transformers for High-Resolution Image Synthesis," *CVPR*, 2021.

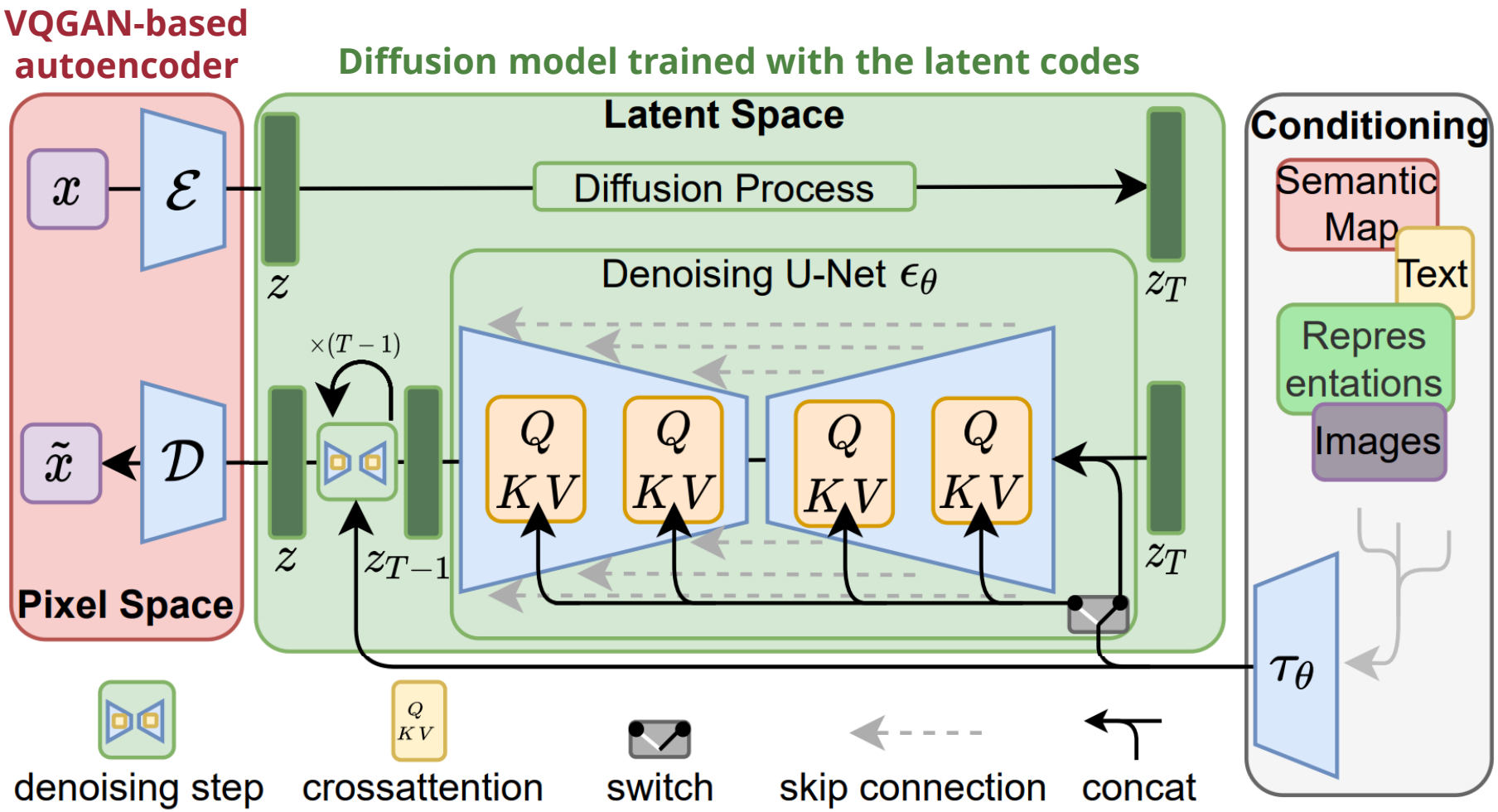
VQGAN: Conditional Generation



(Source: Esser et al., 2021)

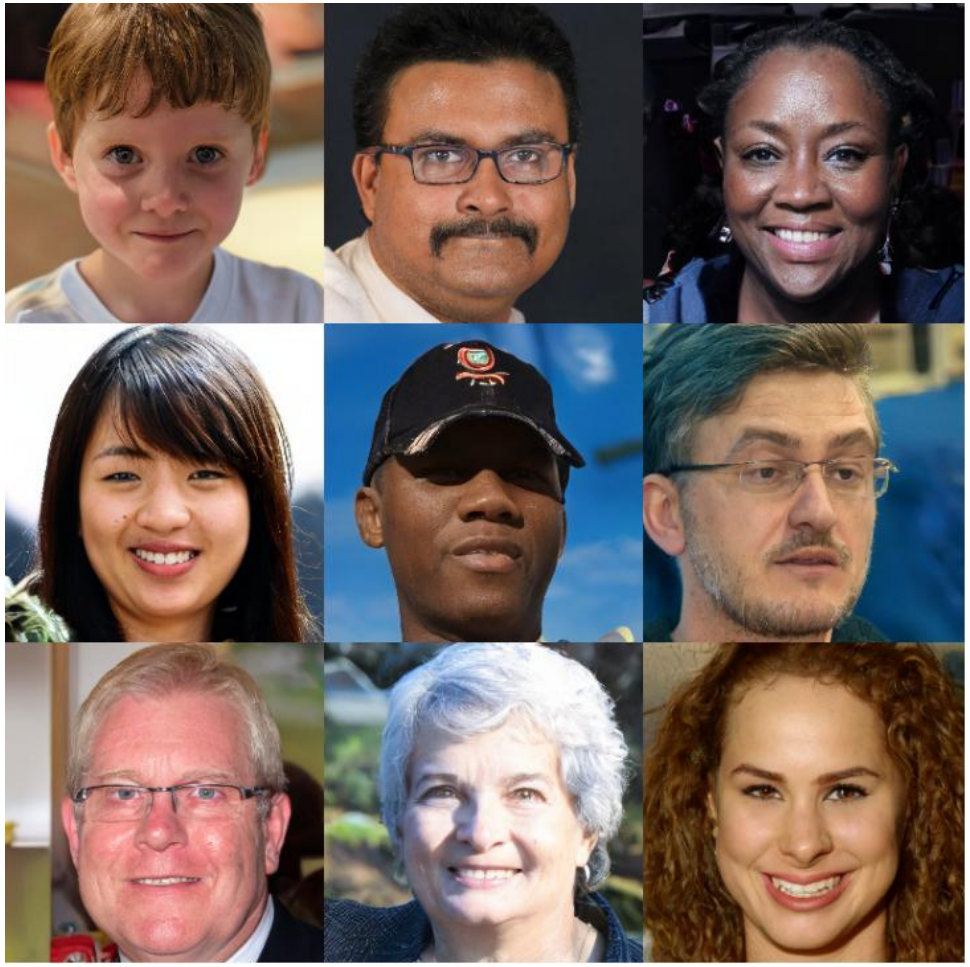
Patrick Esser, Robin Rombach, and Björn Ommer, "Taming Transformers for High-Resolution Image Synthesis," *CVPR*, 2021.

Latent Diffusion Models (LDMs)



(Source: Rombach et al., 2022)

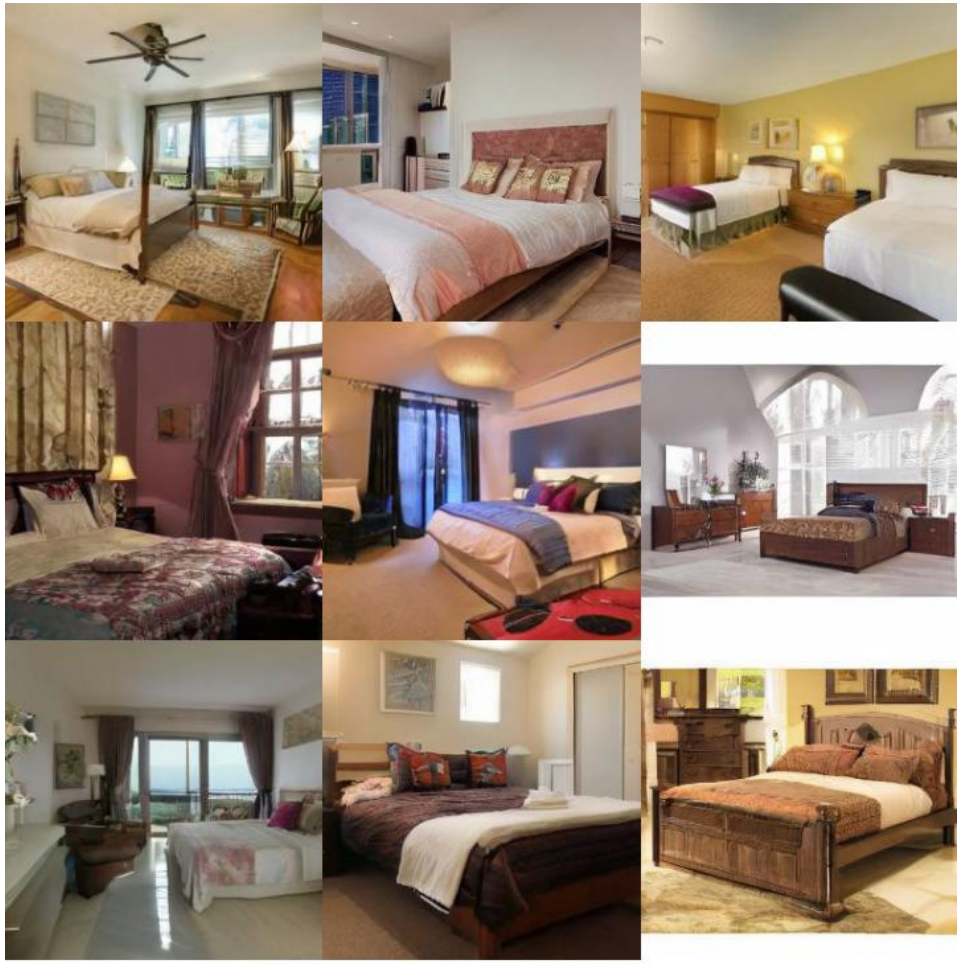
LDMs: Examples



(Source: Rombach et al., 2022)

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *CVPR*, 2022.

LDMs: Examples

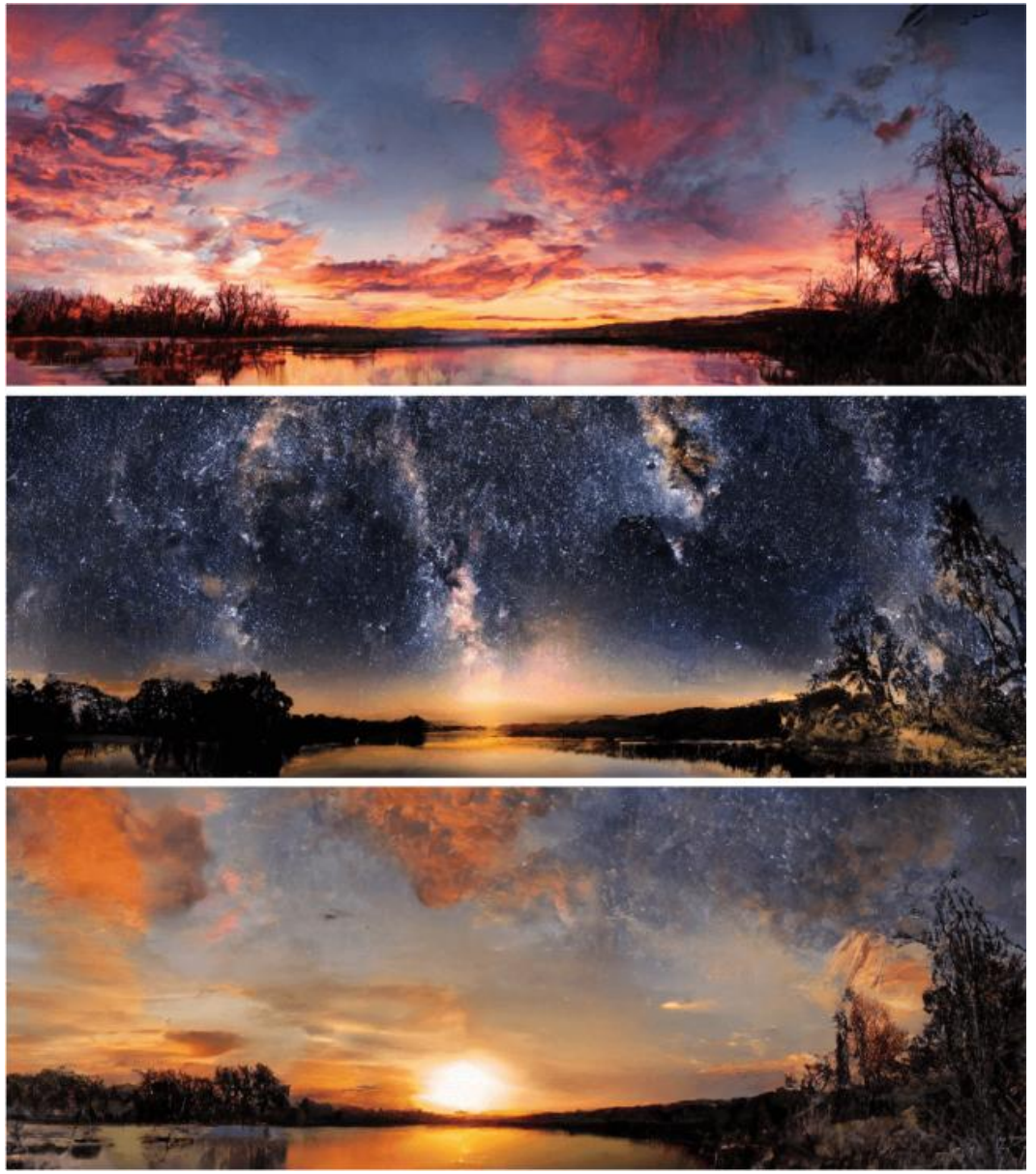


(Source: Rombach et al., 2022)

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *CVPR*, 2022.

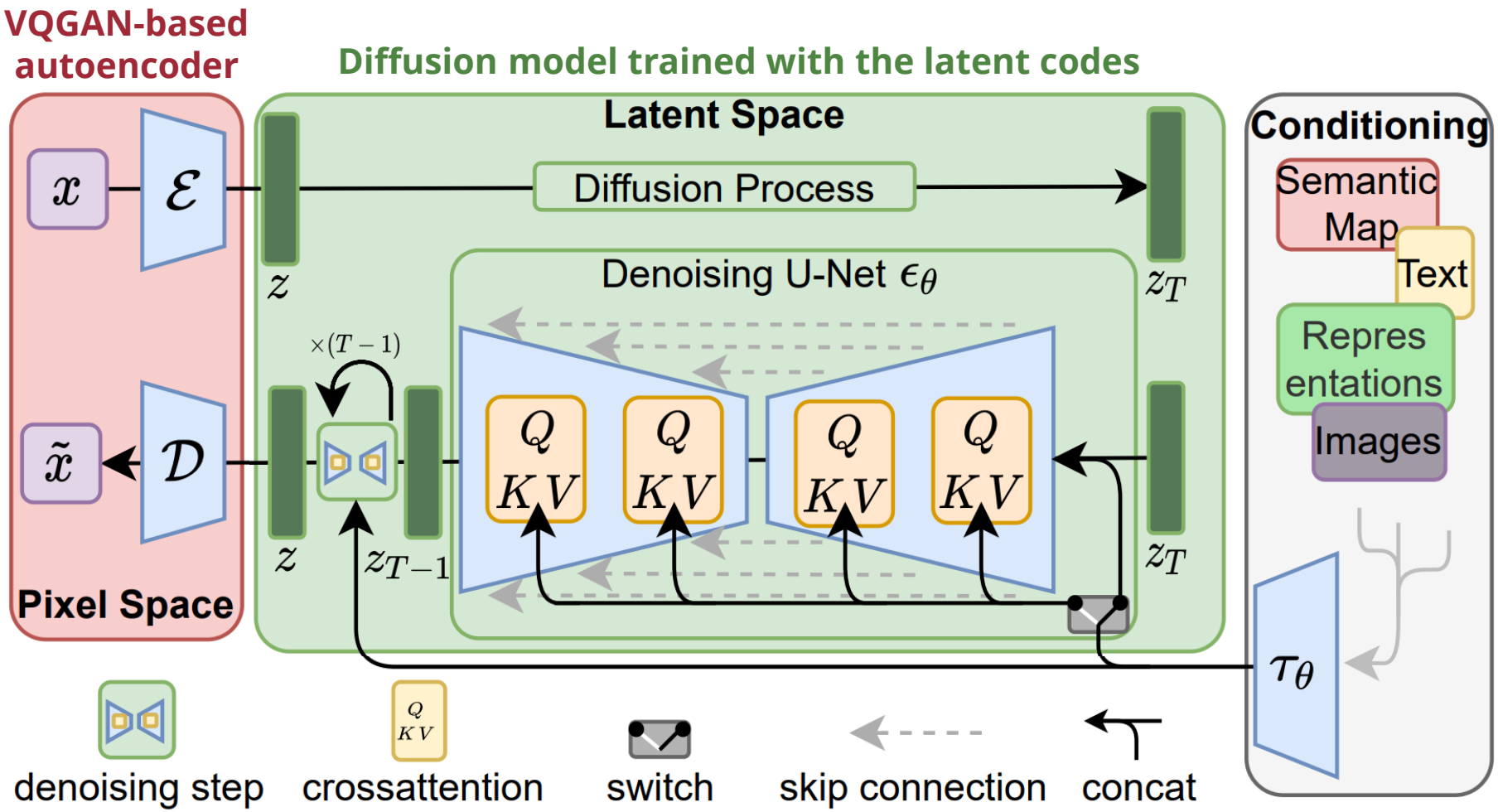
LDMs: Semantic Synthesis

(Source: Rombach et al., 2022)



Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *CVPR*, 2022.

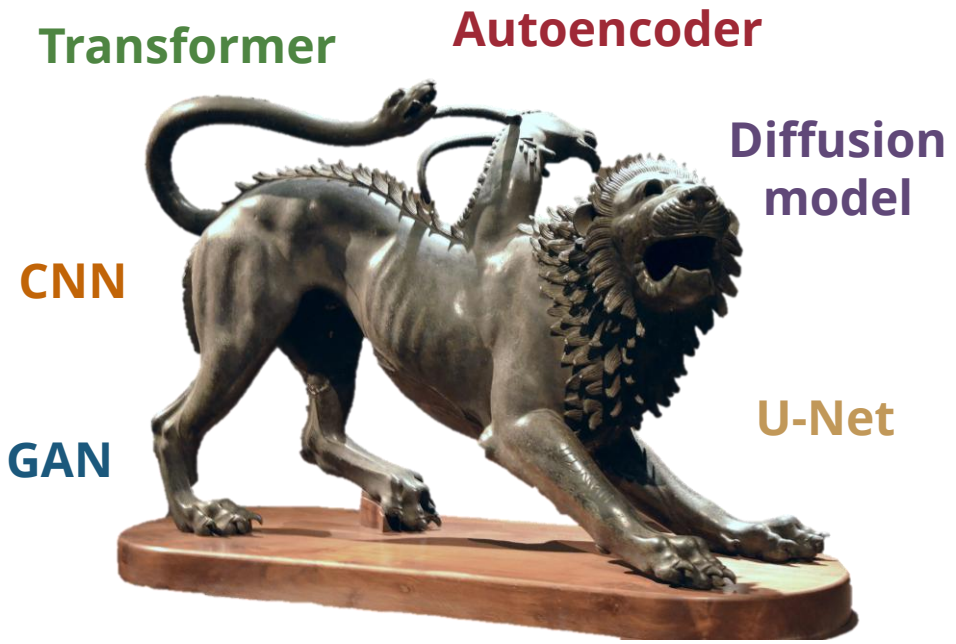
Latent Diffusion Models (LDMs)



(Source: Rombach et al., 2022)

Latent Diffusion Model is a Chimera

- **A neural codec**
  - An CNN-based autoencoder
  - Trained with a GAN-like adversarial loss
- **Diffusion model in the latent space**
  - A denoising U-Net
- **A conditioning module**
  - Transformer-like cross-attention mechanism

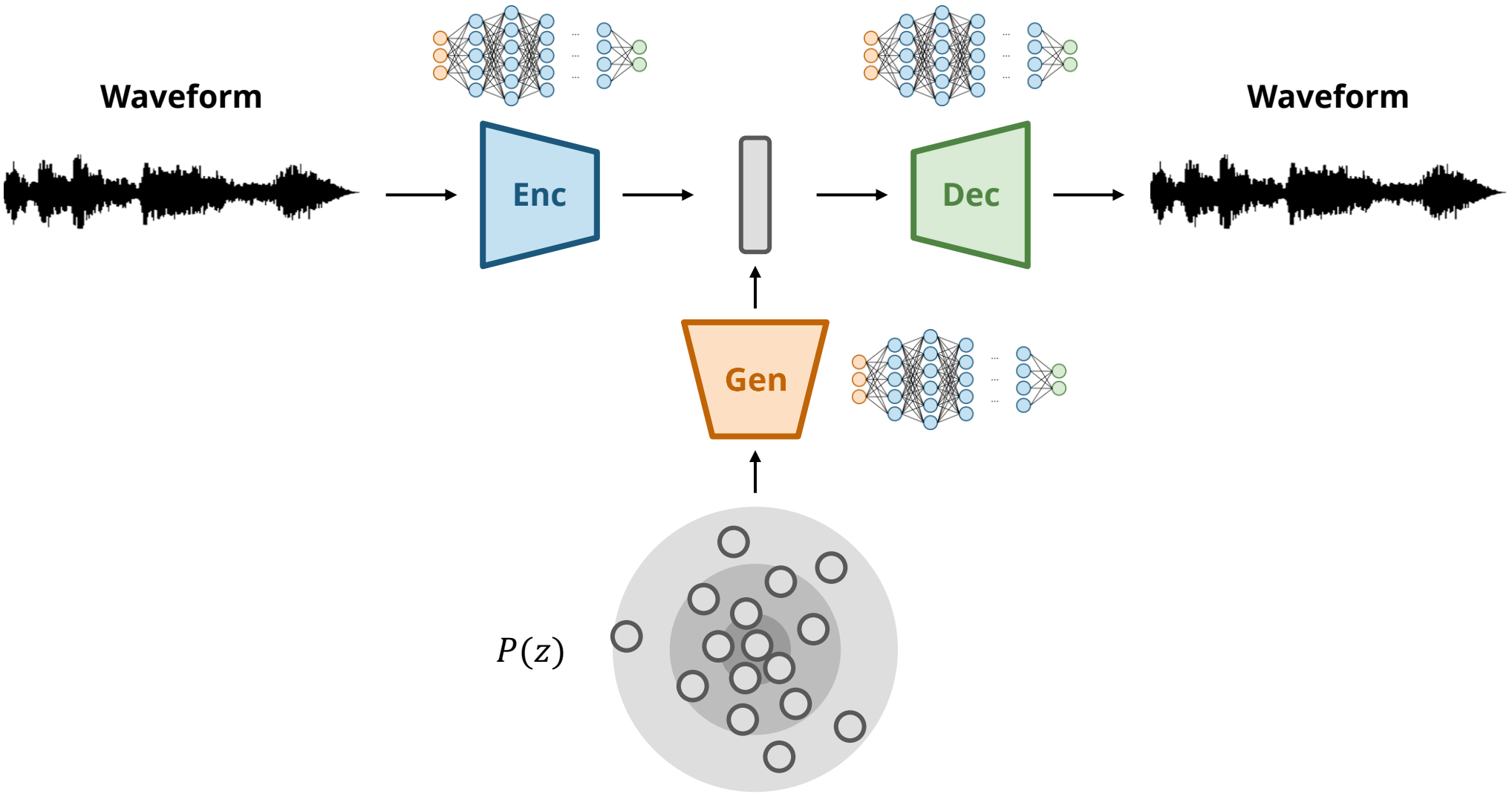


(Source: Raddato via worldhistory.org)

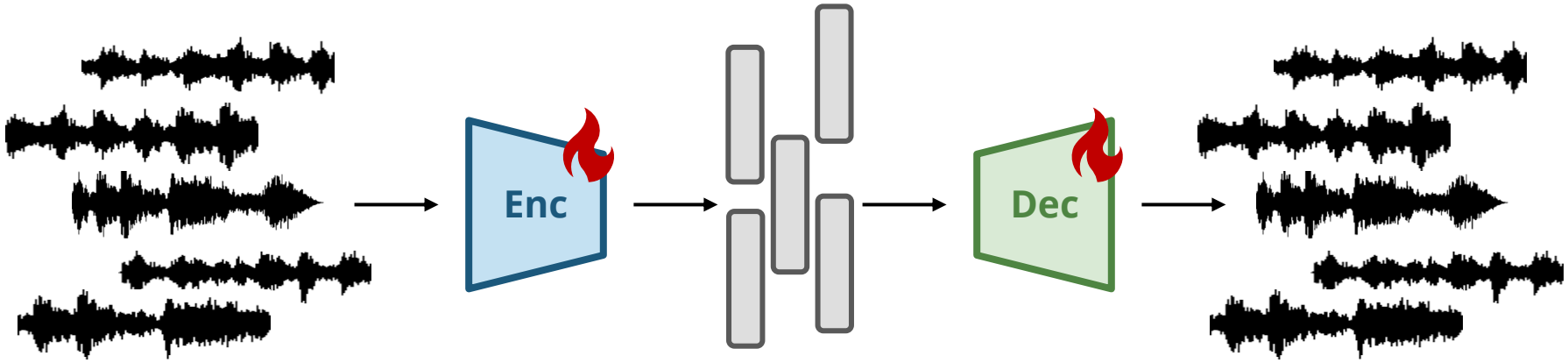
worldhistory.org/image/5462/chimera-of-arezzo-florence/

Latent-based Audio Synthesis

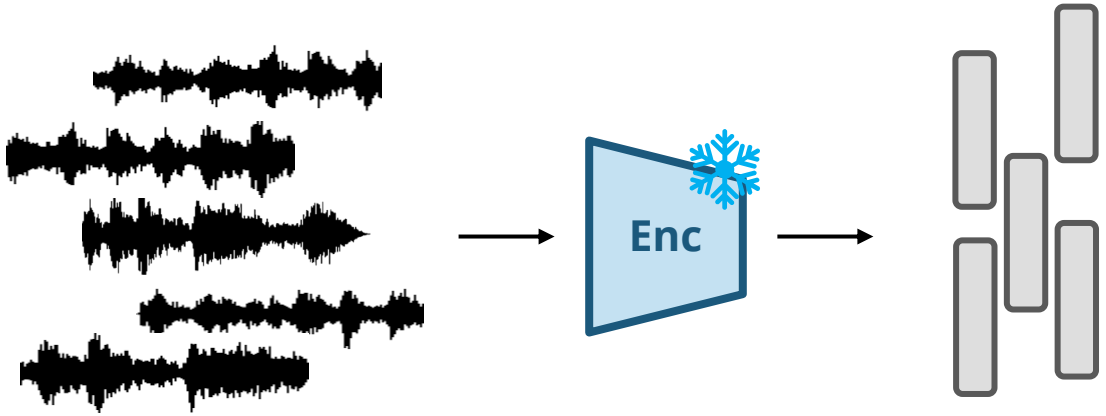
Latent-based Audio Synthesis



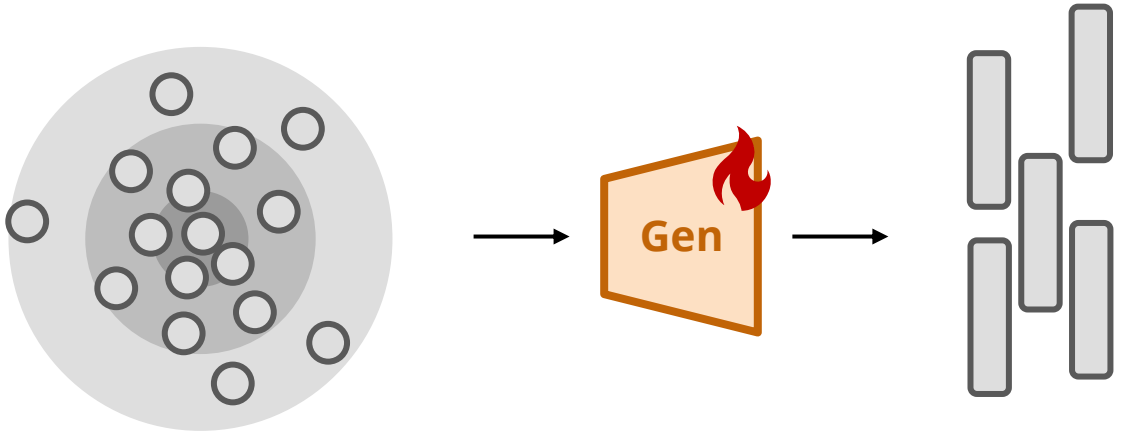
Step 1: Train an Autoencoder



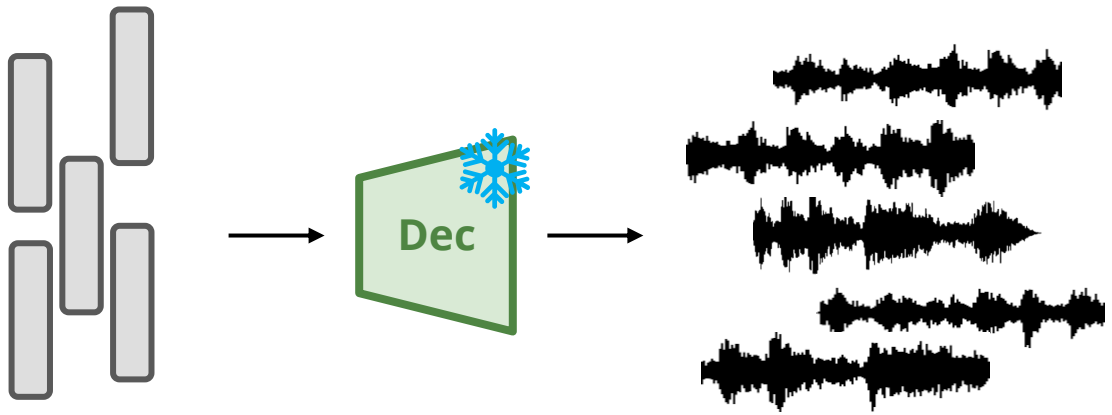
Step 2: Compute the Latent Vectors



Step 3: Train a Latent Generative Model

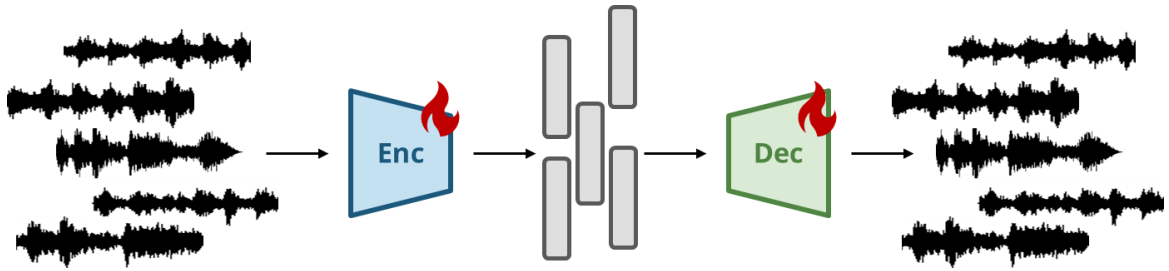


Step 4: Decode the Latent Vectors

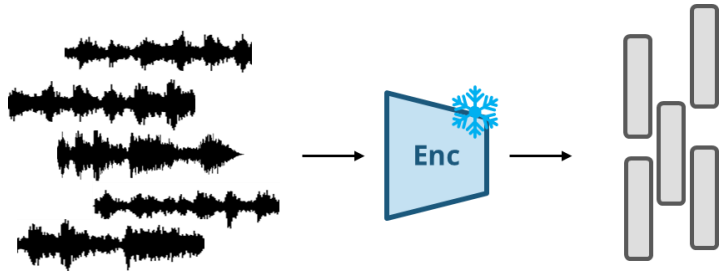


Pipeline

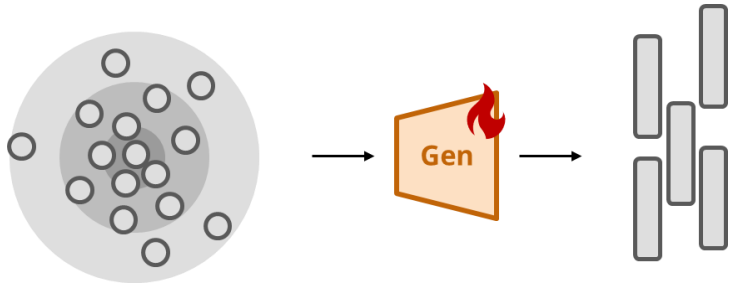
Step 1: Train an Autoencoder



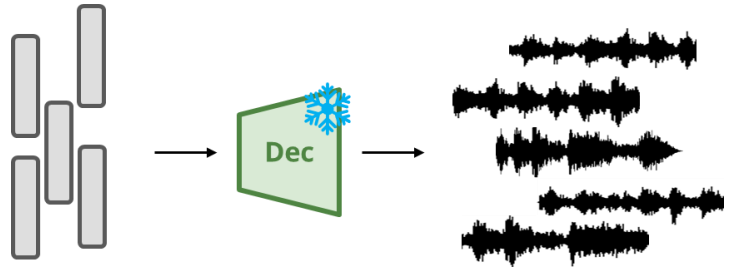
Step 2: Compute the Latent Vectors



Step 3: Train a Latent Generative Model

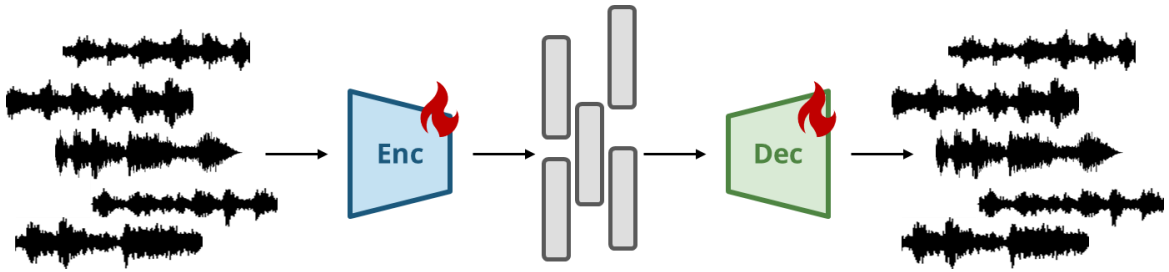


Step 4: Decode the Latent Vectors

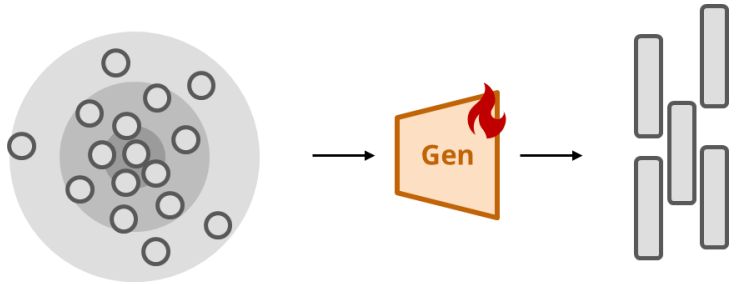


Training

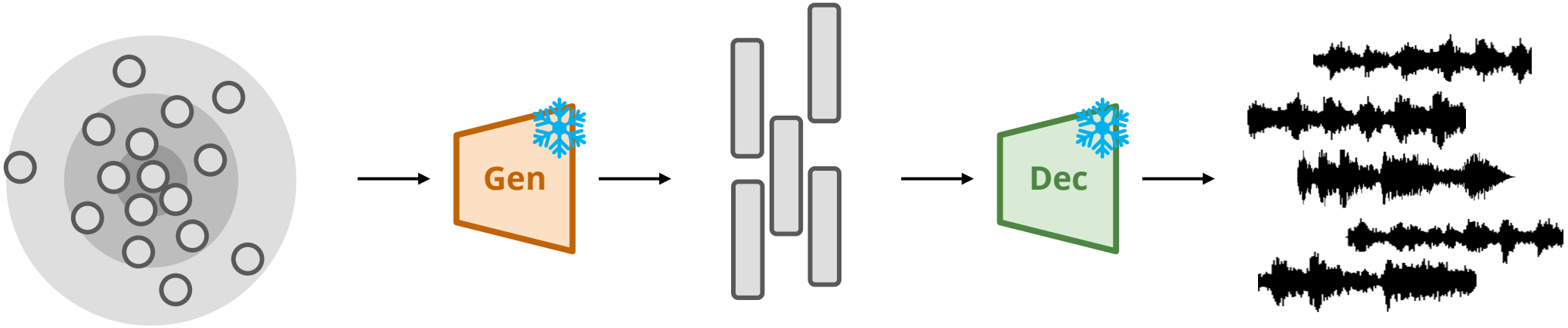
Autoencoder



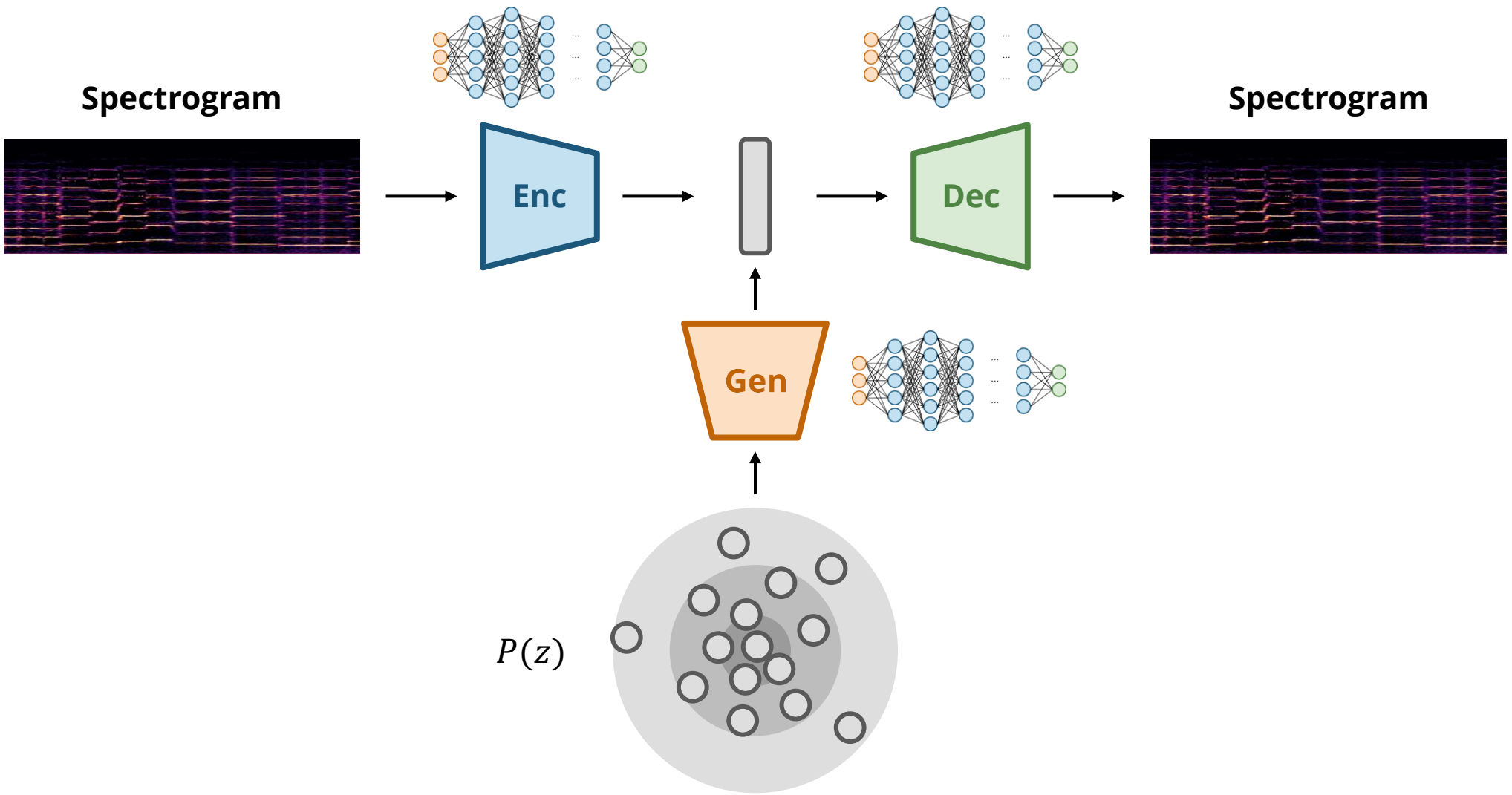
Latent Generative Model



Inference



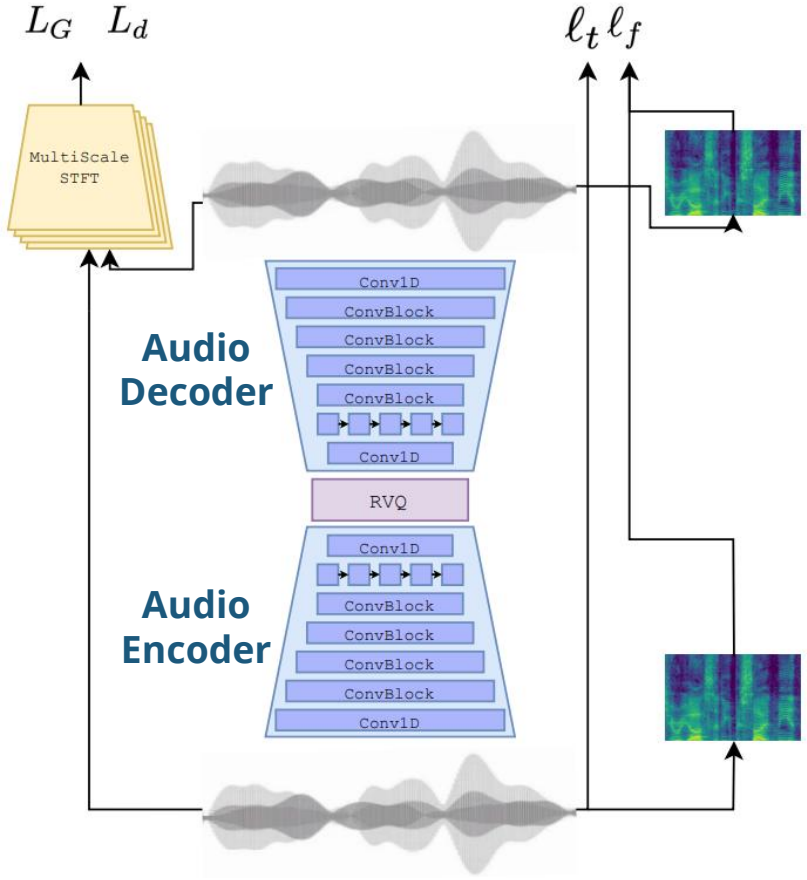
Latent-based Audio Synthesis



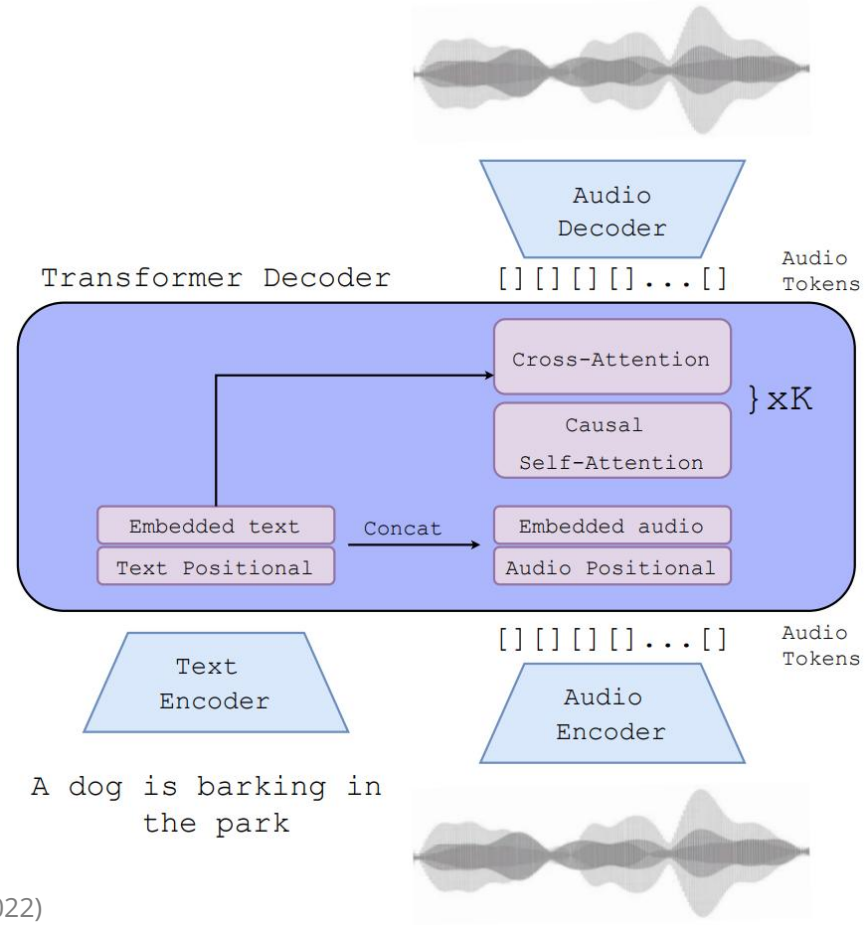
Latent-based Audio Synthesis

AudioGen (Kreuk et al., 2023)

Audio Autoencoder



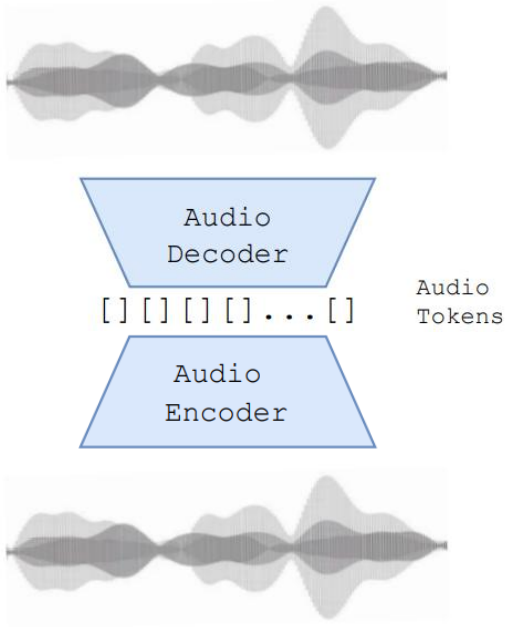
Audio Language Model



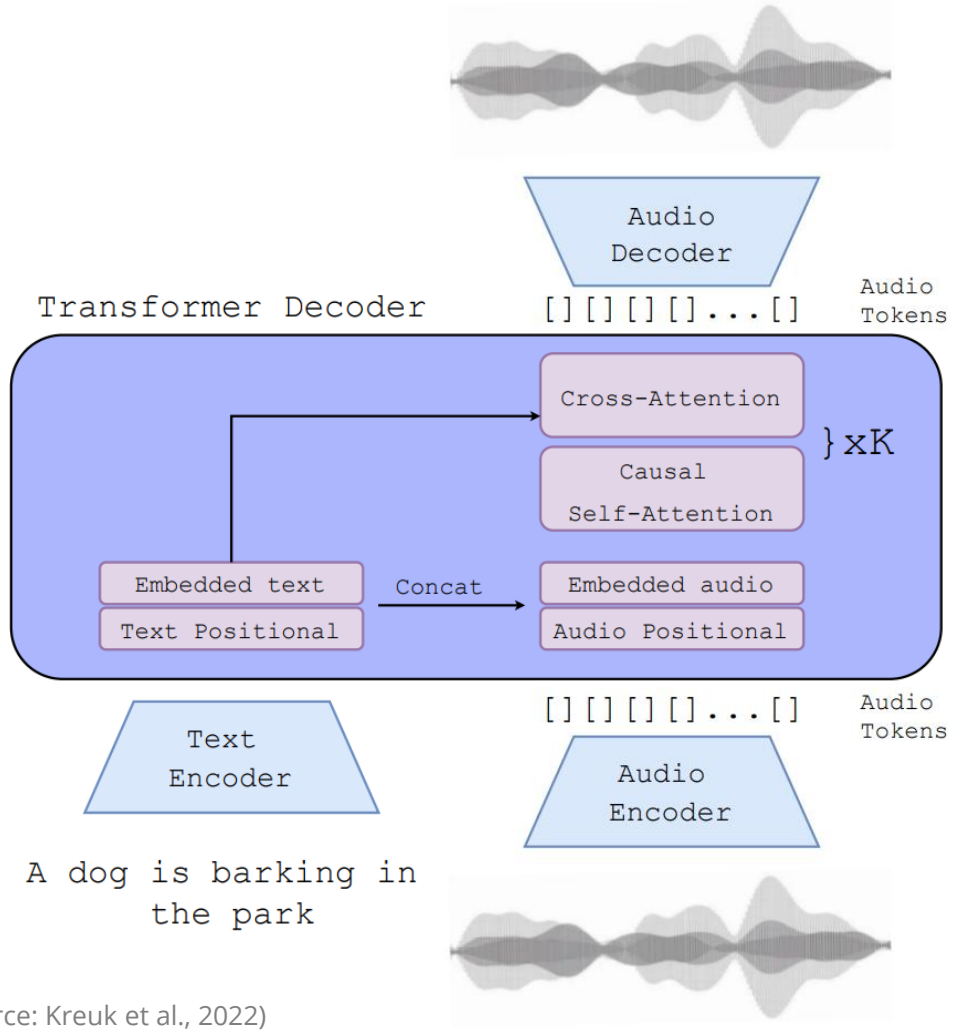
(Source: Kreuk et al., 2022)

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "AudioGen: Textually Guided Audio Generation," *ICLR*, 2023.

AudioGen (Kreuk et al., 2023)



4k hours
(speech, music, sound effects)



(Source: Kreuk et al., 2022)

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "AudioGen: Textually Guided Audio Generation," *ICLR*, 2023.

AudioGen: Examples (Kreuk et al., 2023)

(Source: Kreuk et al., 2022)

felixkreuk.github.io/audiogen

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "AudioGen: Textually Guided Audio Generation," *ICLR*, 2023.

MusicGen (Copet et al., 2023)

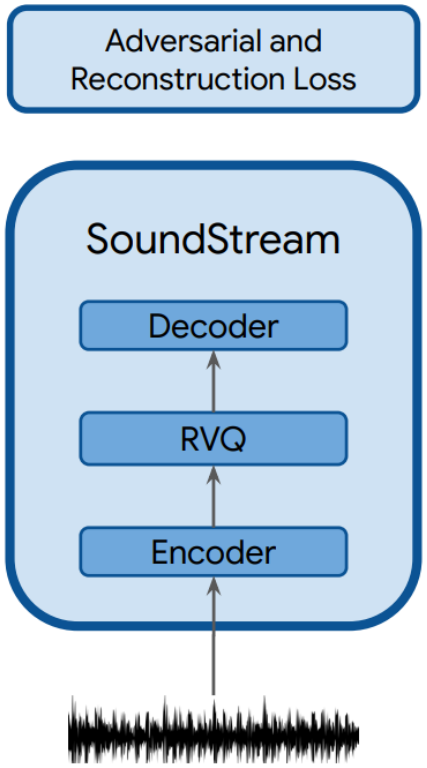
- AudioGen for Music
- Use EnCodec (Défossez et al., 2022) as the autoencoder
  - instead of SoundStream for AudioGen (Kreuk et al., 2023)
- **20k hours** of licensed music
  - Internal dataset 10k High-quality (private)
  - ShutterStock 25k Instrument-only
  - Pond5 365k Instrument-only

ai.honu.io/papers/musicgen/

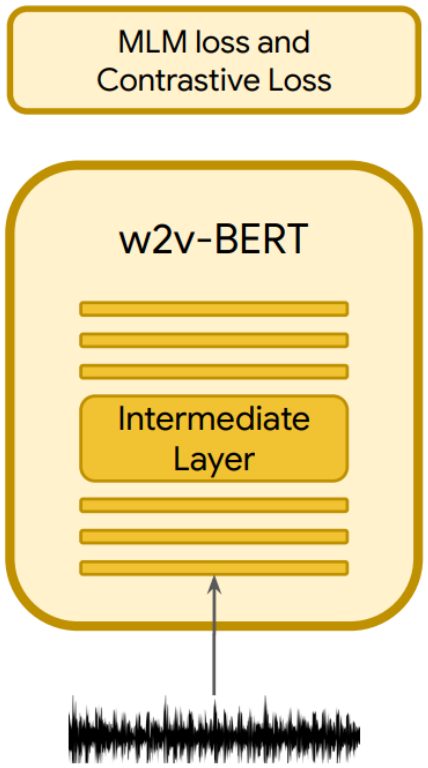
Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and Controllable Music Generation," *NeurIPS*, 2023.

MusicLM (Agostinelli et al., 2023)

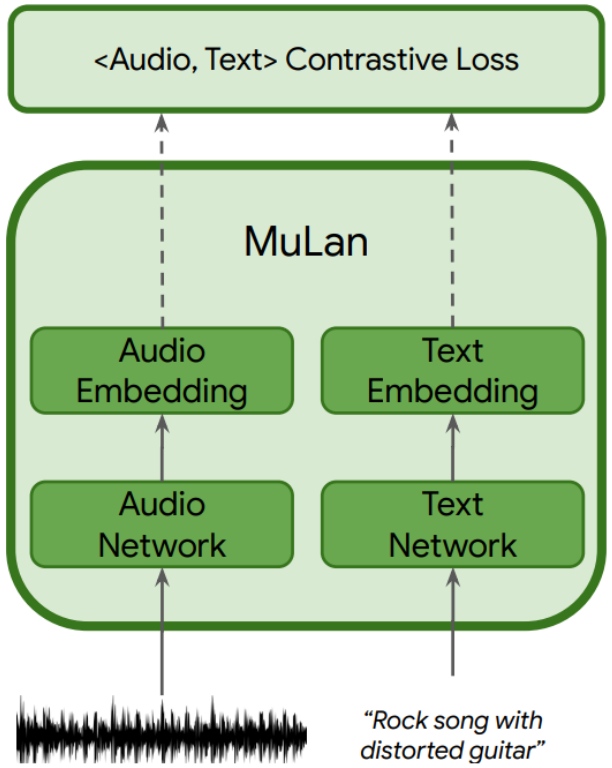
Audio autoencoder



Semantic representation



Text-music correspondence



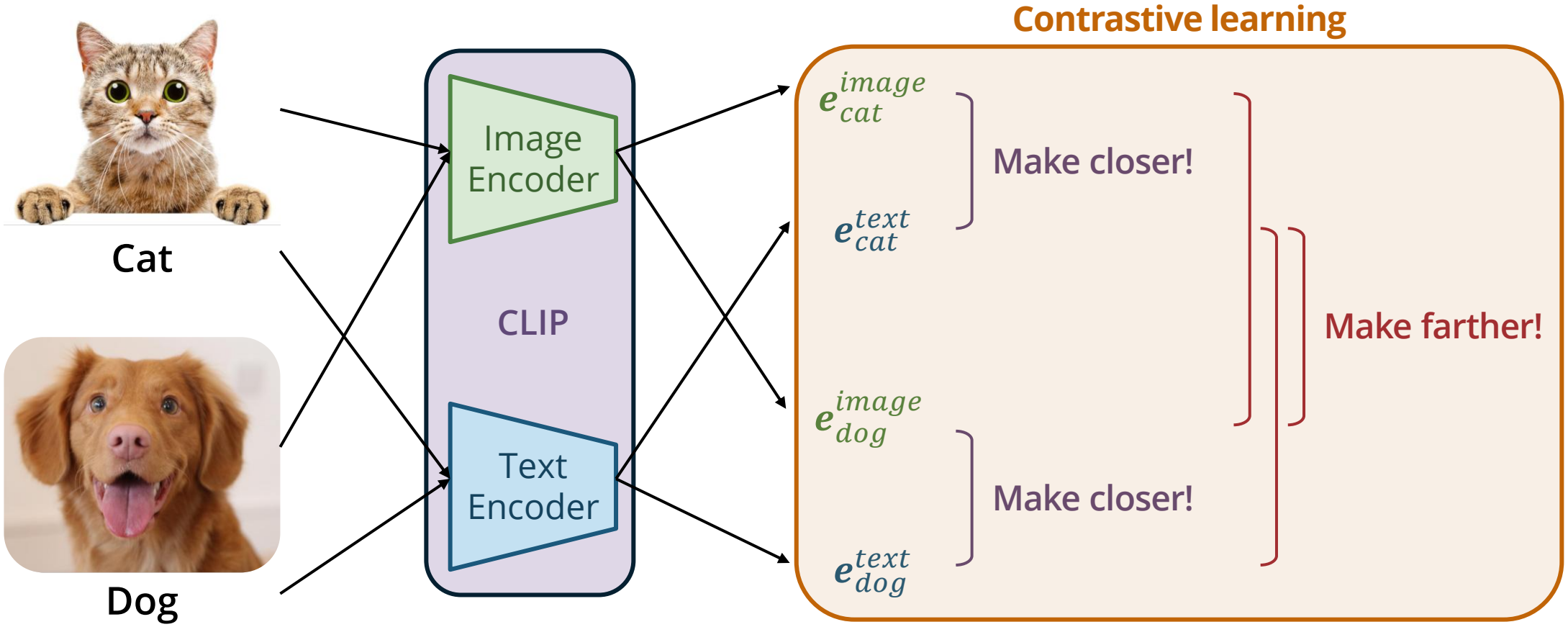
106k songs, 8.2k hours

44M 30-sec clips, 370k hours

(Source: Agostinelli et al., 2022)

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 2023.

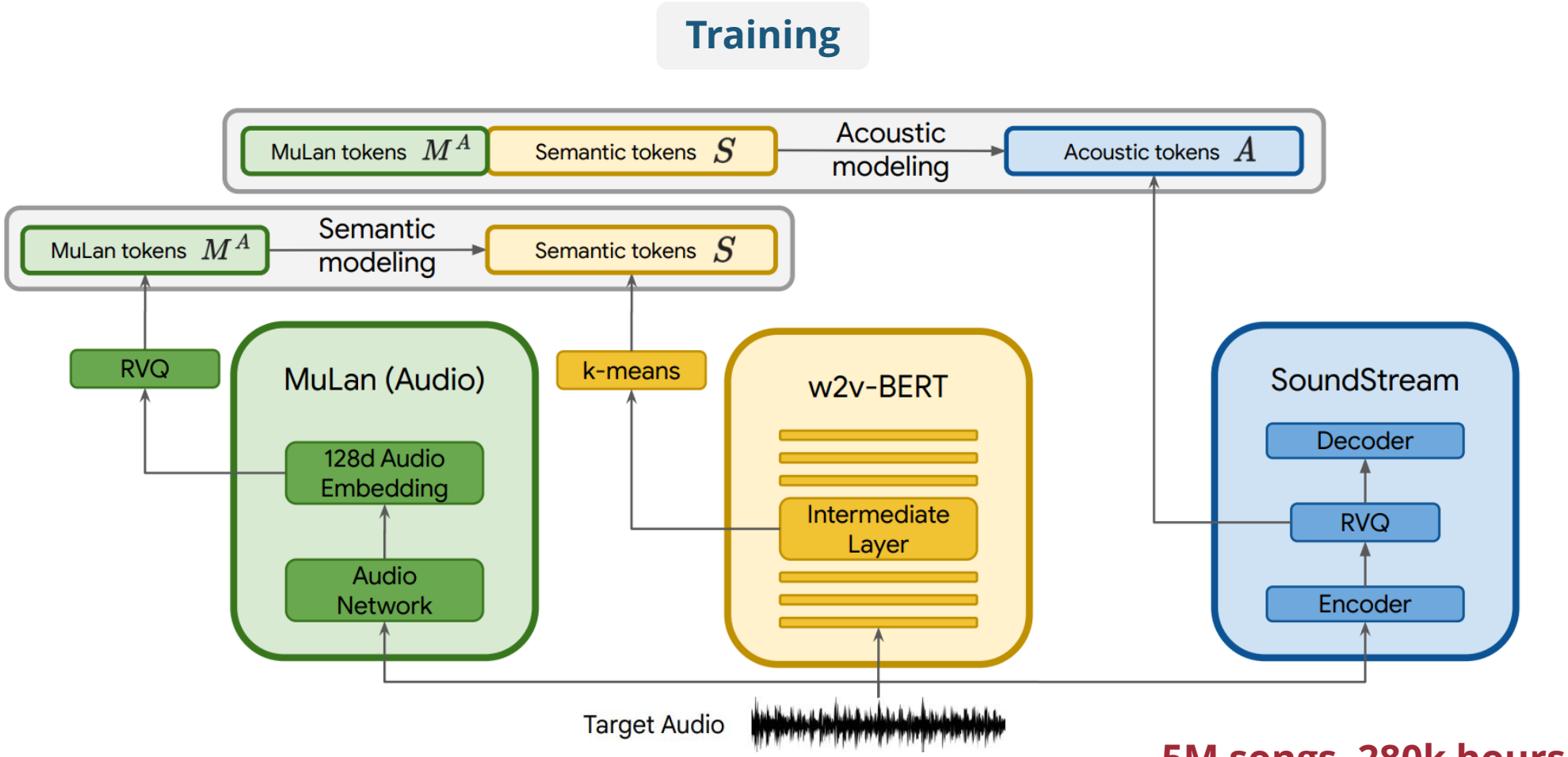
Contrastive Language-Image Pretraining (CLIP)



Learn a shared embedding space for images and texts

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *ICML*, 2021.

MusicLM (Agostinelli et al., 2023)

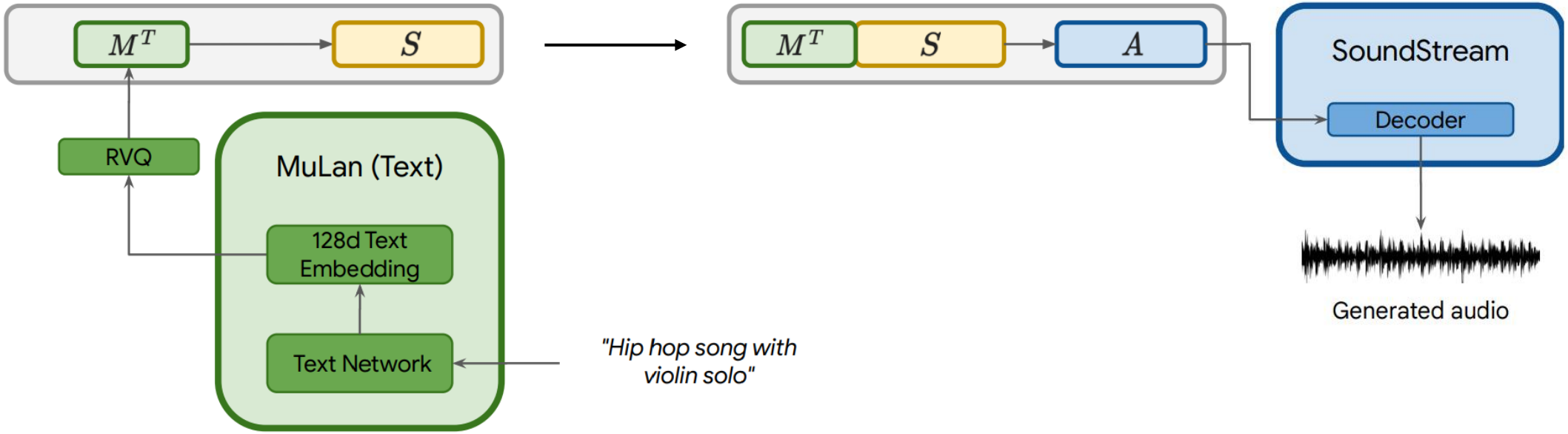


(Source: Agostinelli et al., 2022)

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 2023.

MusicLM (Agostinelli et al., 2023)

Inference



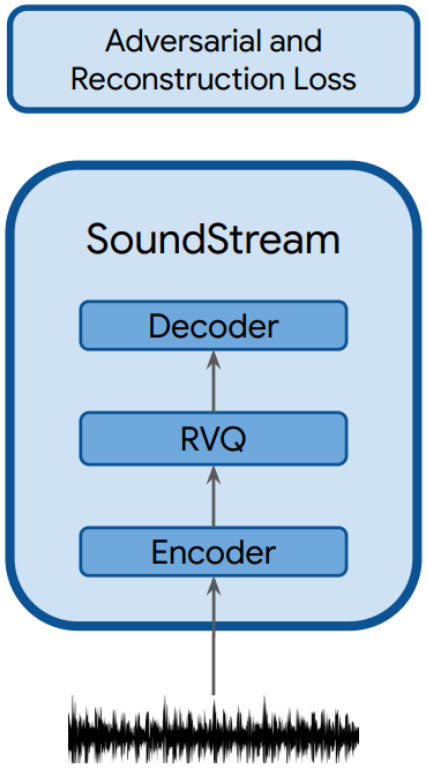
(Source: Agostinelli et al., 2022)

google-research.github.io/seanet/musiclm/examples/

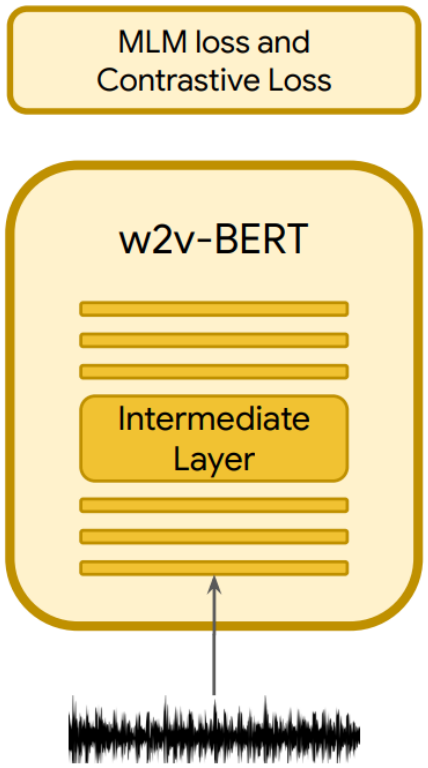
Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 2023.

MusicLM (Agostinelli et al., 2023)

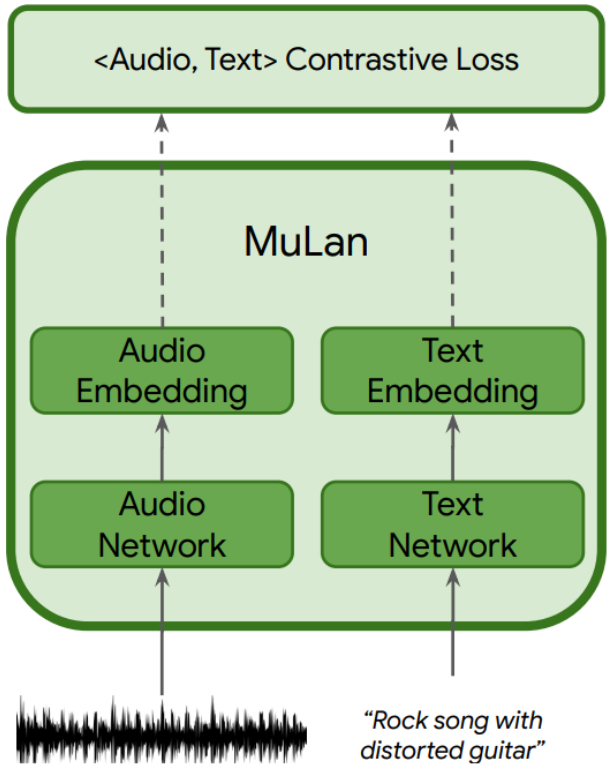
Audio autoencoder



Semantic representation



Text-music correspondence



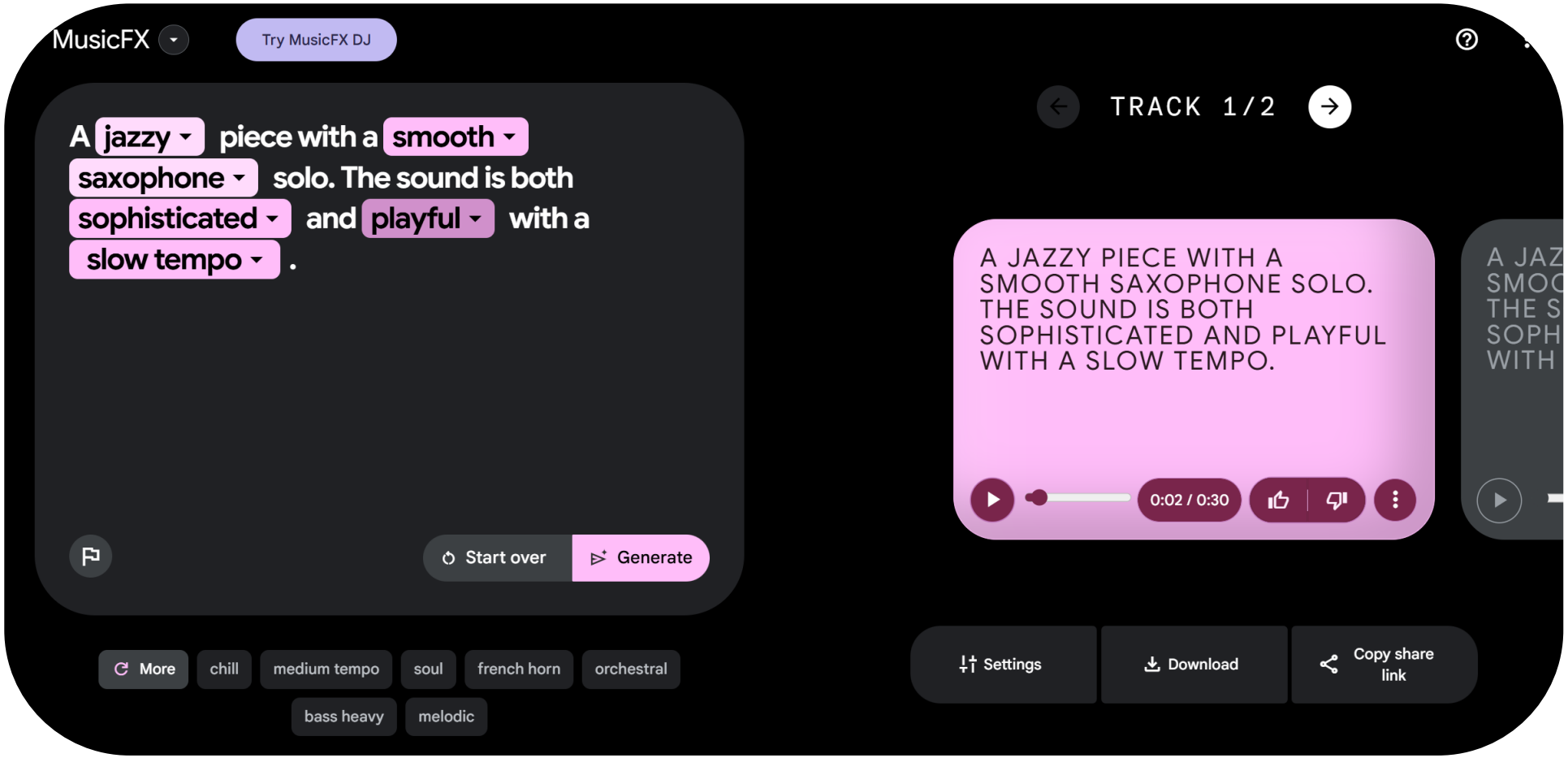
106k songs, 8.2k hours

44M 30-sec clips, 370k hours

(Source: Agostinelli et al., 2022)

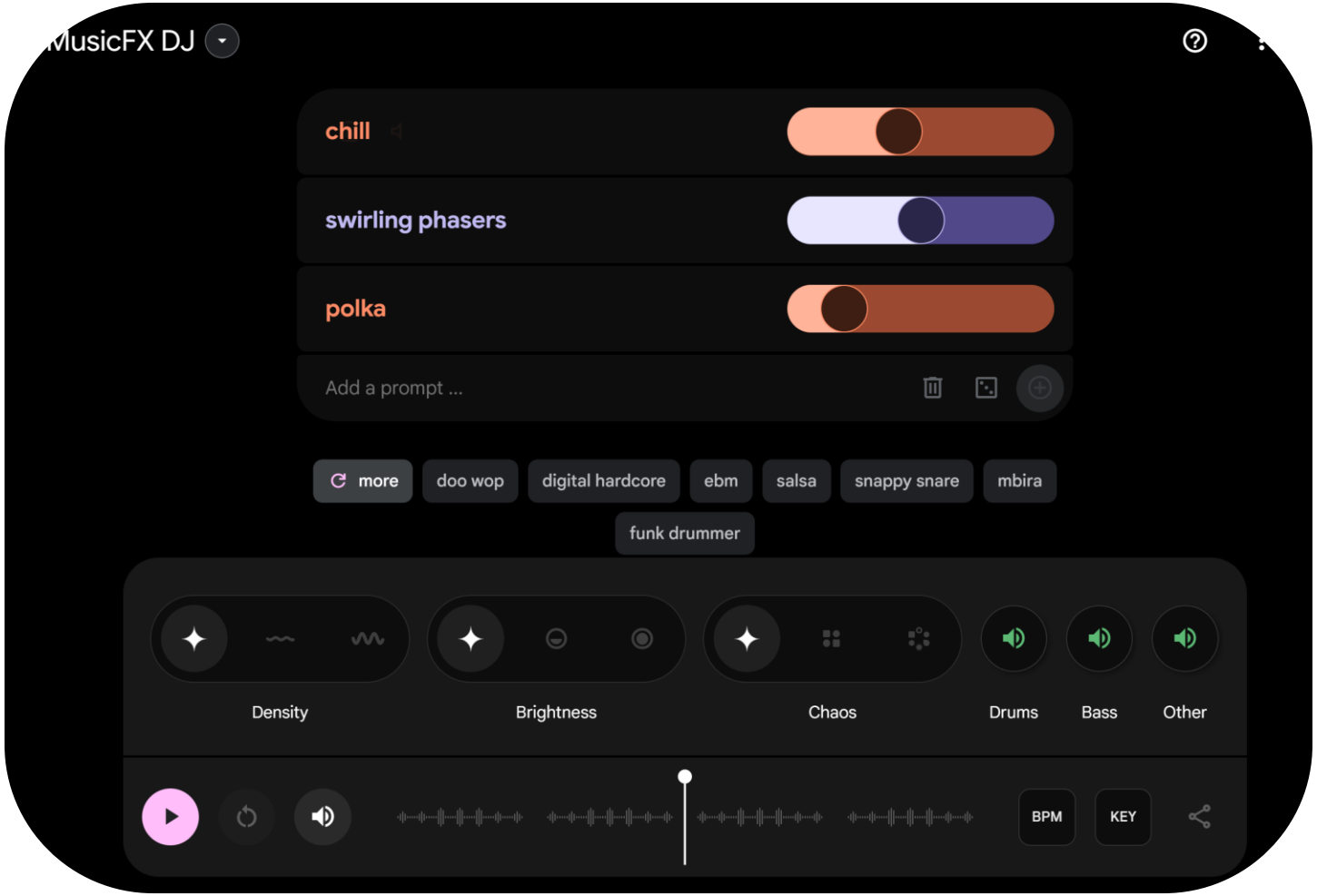
Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 2023.

Music FX (2024)



aitestkitchen.withgoogle.com/tools/music-fx

Music FX DJ (2024)

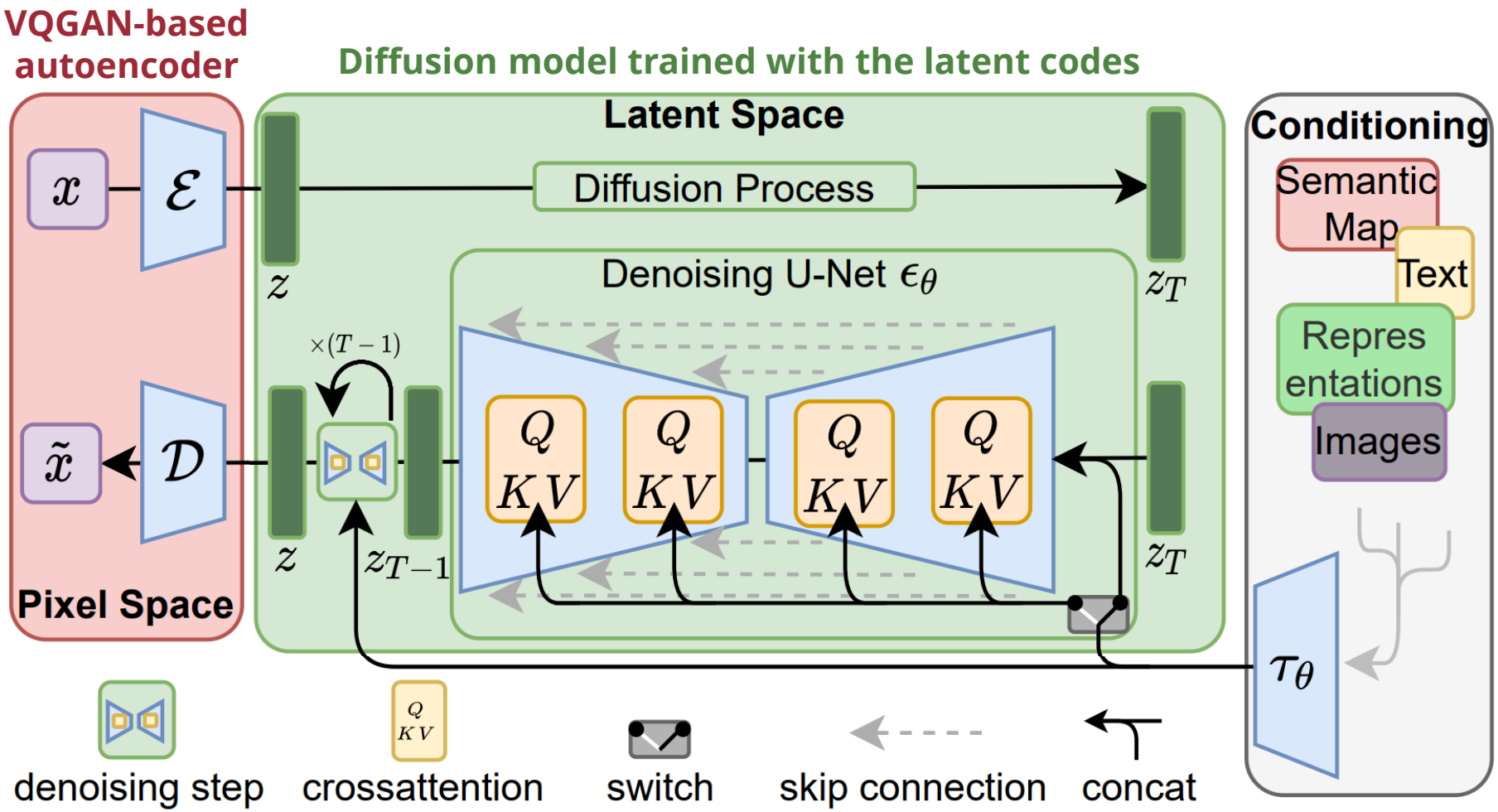


aitestkitchen.withgoogle.com/tools/music-fx-dj

Music FX DJ (2024)

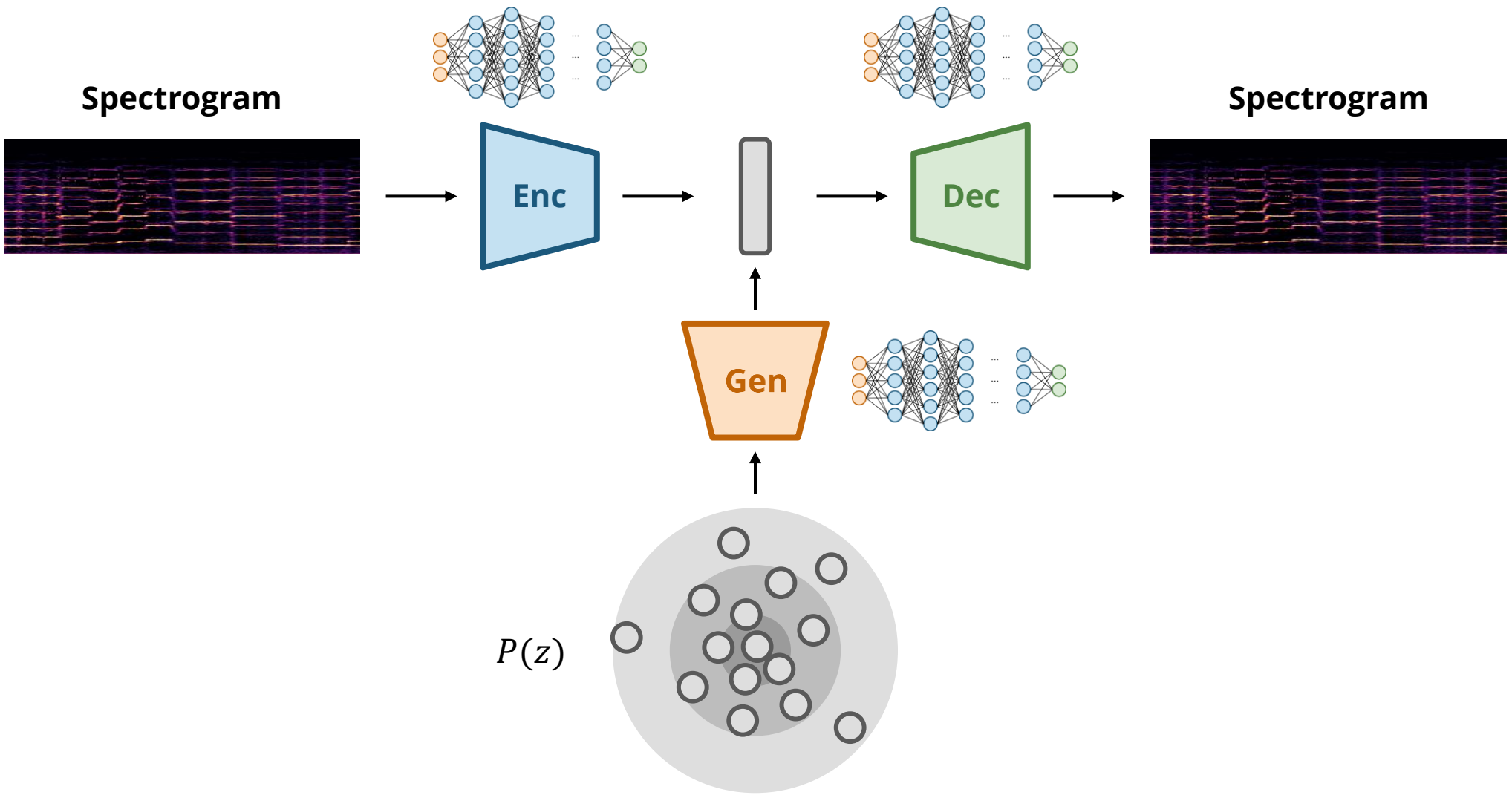
youtube.com/live/IUQW5LgBZvQ

(Recap) Latent Diffusion Models (LDMs)

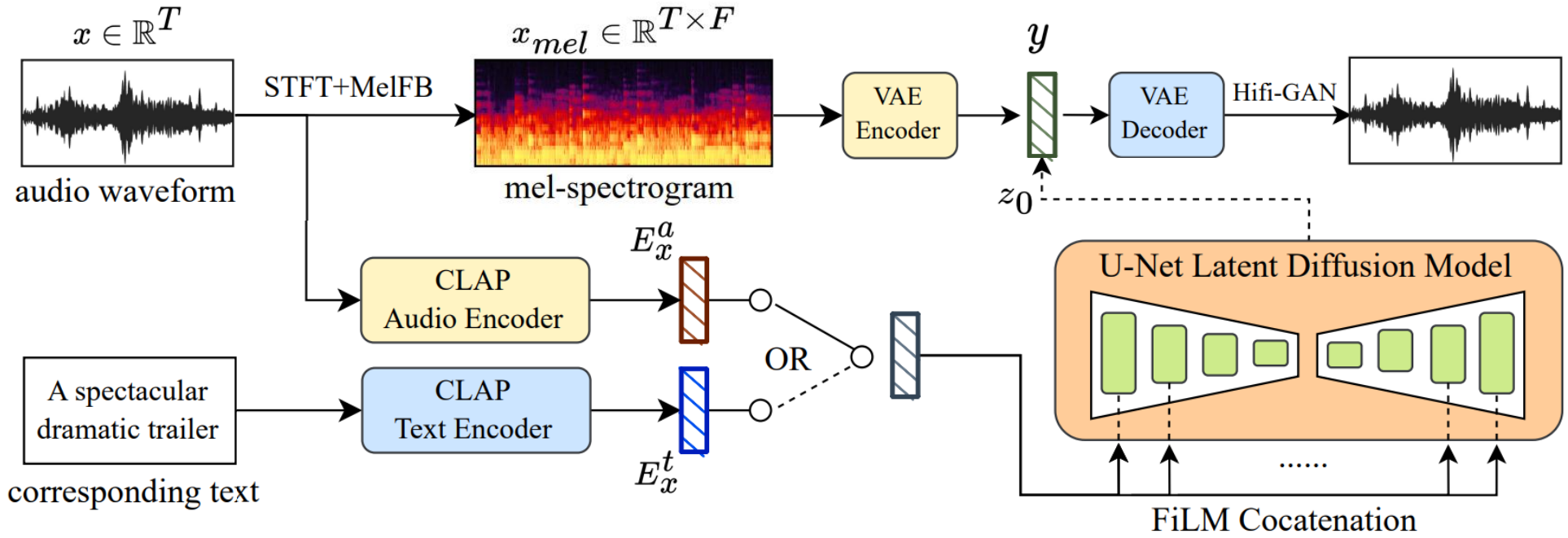


(Source: Rombach et al., 2022)

(Recap) Latent-based Audio Synthesis



Example: MusicLDM (Chen et al., 2023)

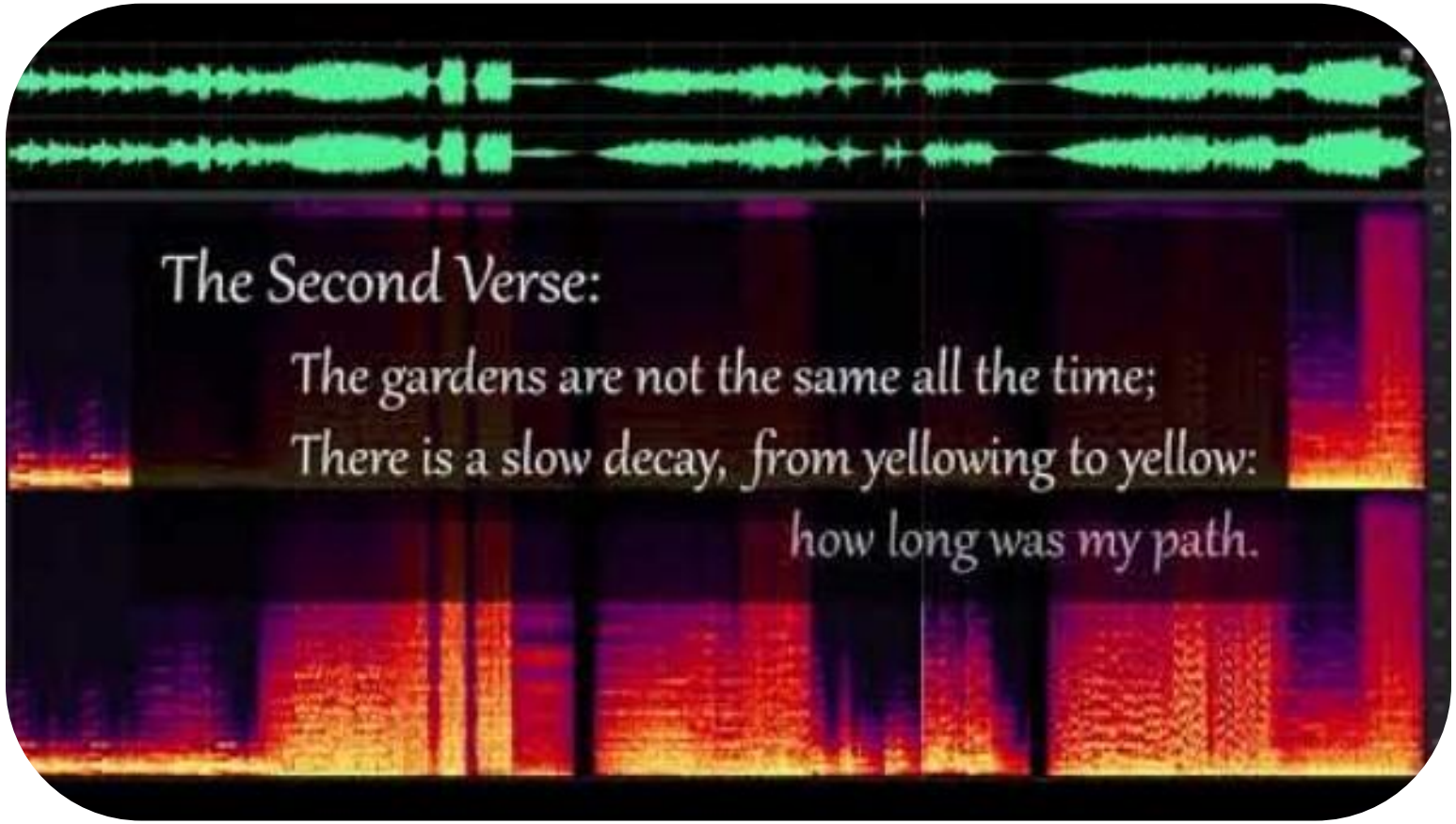


(Source: Ke et al., 2023)

musicldm.github.io

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies," *ICASSP*, 2024.

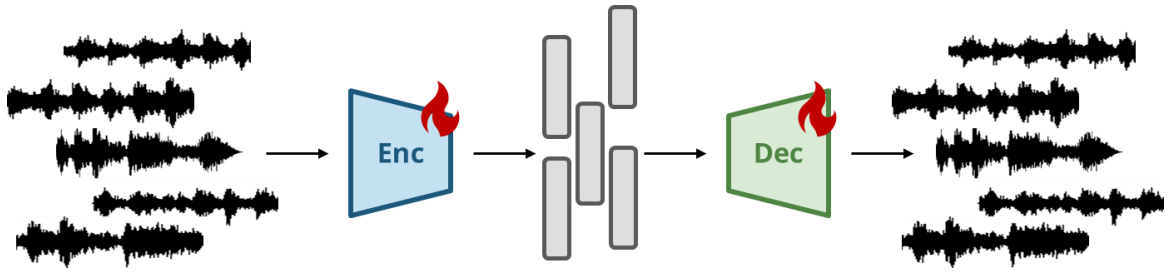
Example: MusicLDM (Chen et al., 2023)



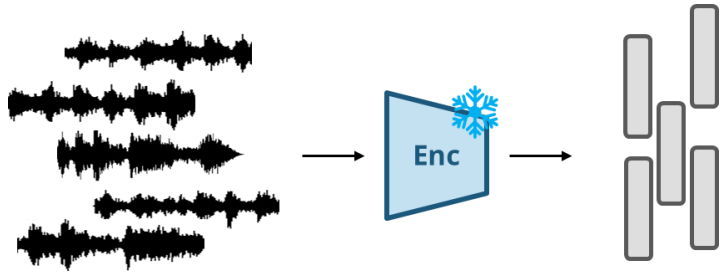
youtu.be/DALv7ea6cv0

Pipeline

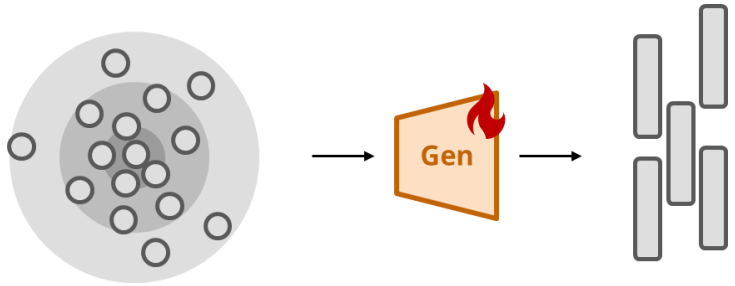
Step 1: Train an Autoencoder



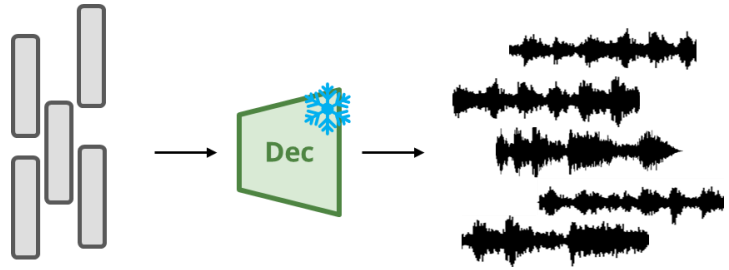
Step 2: Compute the Latent Vectors



Step 3: Train a Latent Generative Model



Step 4: Decode the Latent Vectors



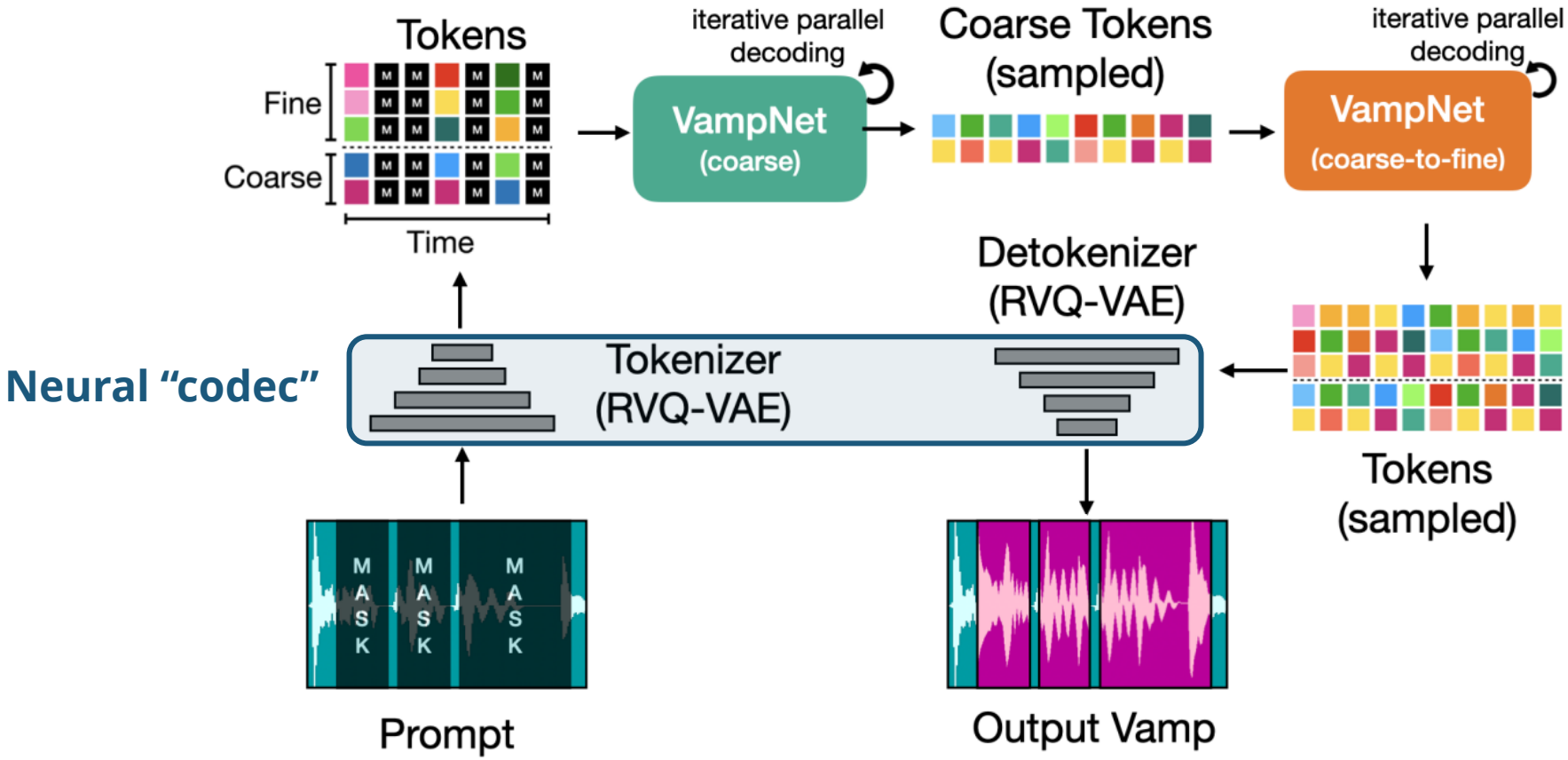
Creative Applications of Music Generation Systems

`unloop` (Garcia et al., 2023)

youtu.be/yzBI8Vcjd2s

github.com/hugofloresgarcia/unloop

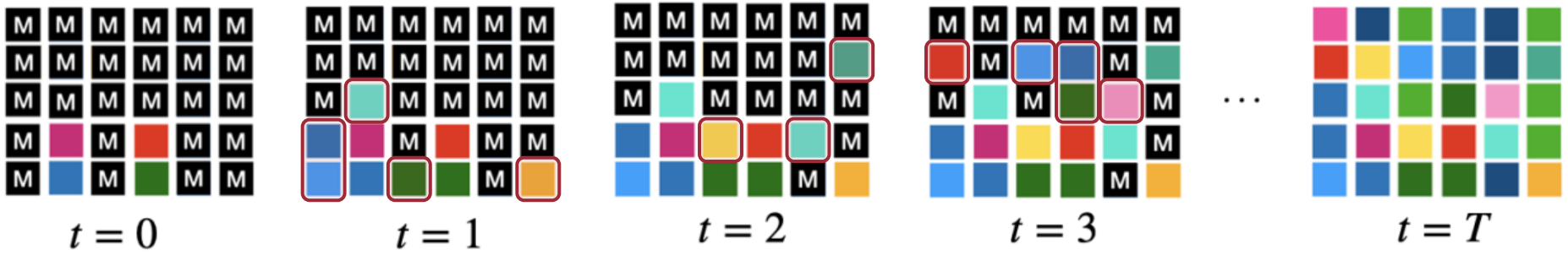
VampNet (Garcia et al., 2023)



(Source: Garcia et al., 2023)

Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo, "VampNet: Music Generation via Masked Acoustic Token Modeling," *ISMIR*, 2023.

VampNet (Garcia et al., 2023)

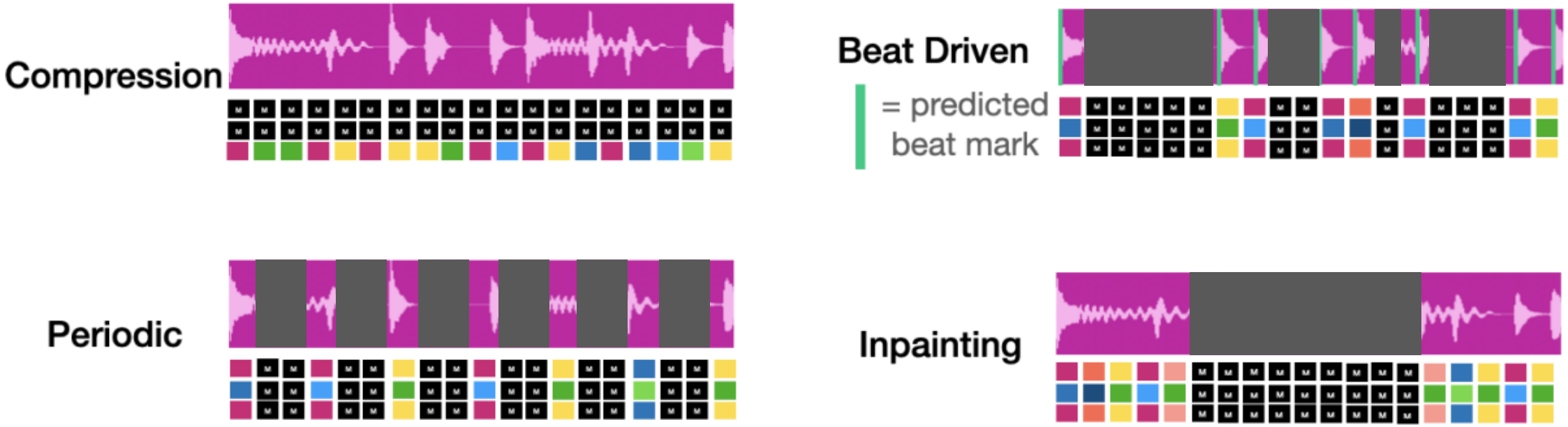


(Source: Garcia et al., 2023)

Sample a subset of the most confident predicted tokens in each iteration

Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo, "VampNet: Music Generation via Masked Acoustic Token Modeling," *ISMIR*, 2023.

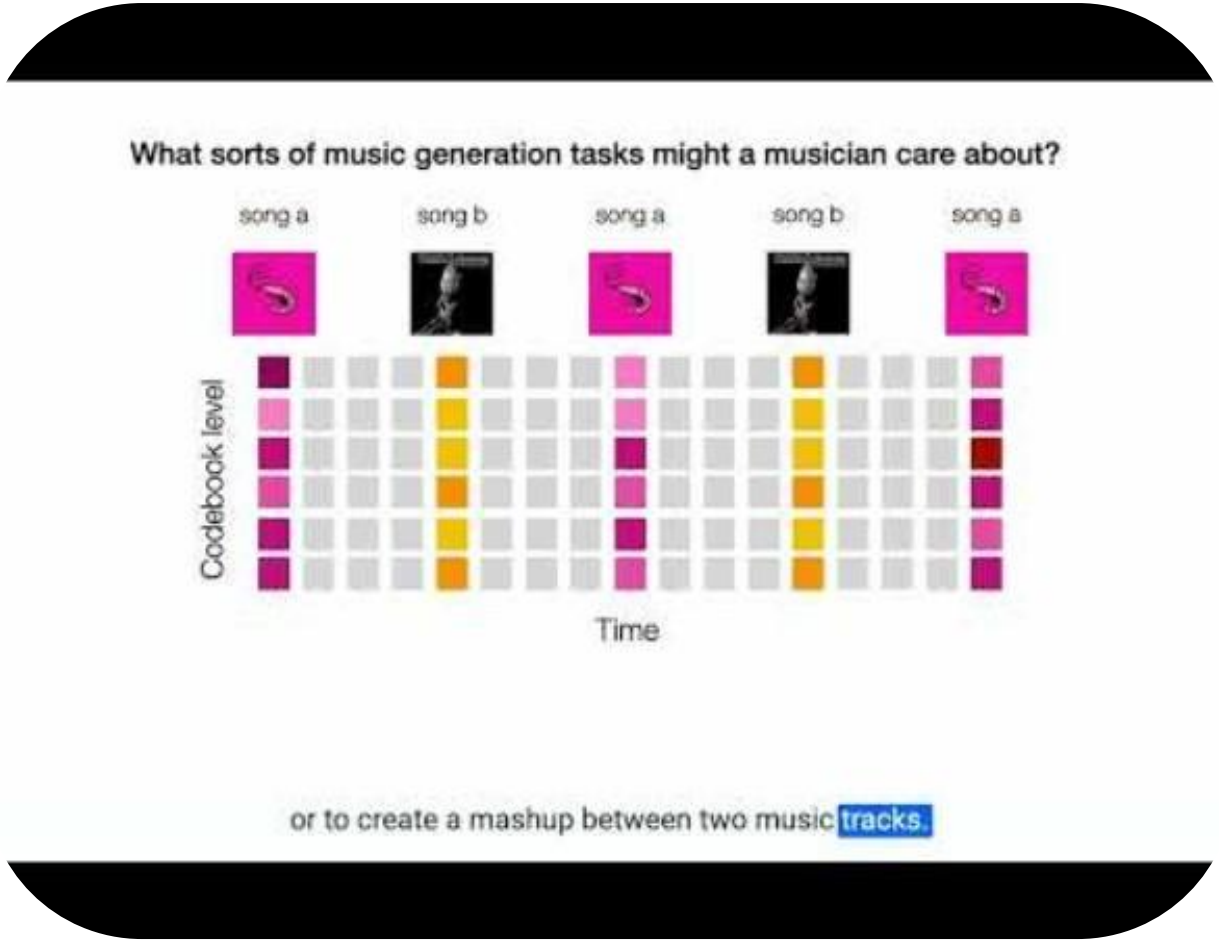
VampNet (Garcia et al., 2023)



(Source: Garcia et al., 2023)

Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo, "VampNet: Music Generation via Masked Acoustic Token Modeling," *ISMIR*, 2023.

VampNet (Garcia et al., 2023)

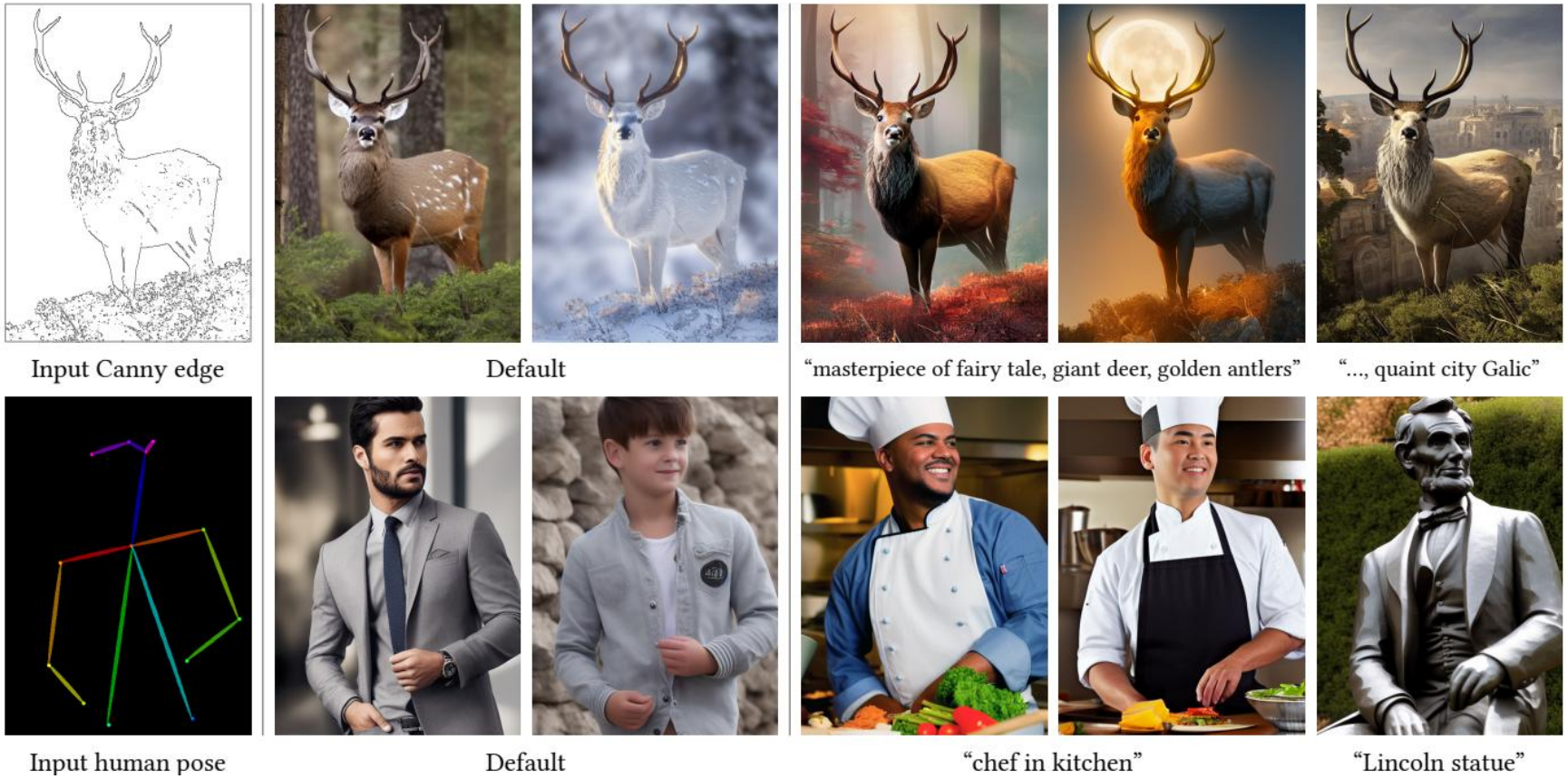


youtu.be/3XfeWlV9Cp0

Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo, "VampNet: Music Generation via Masked Acoustic Token Modeling," *ISMIR*, 2023.

Controlling Music Generation Systems

ControlNet (Zhang et al., 2023)

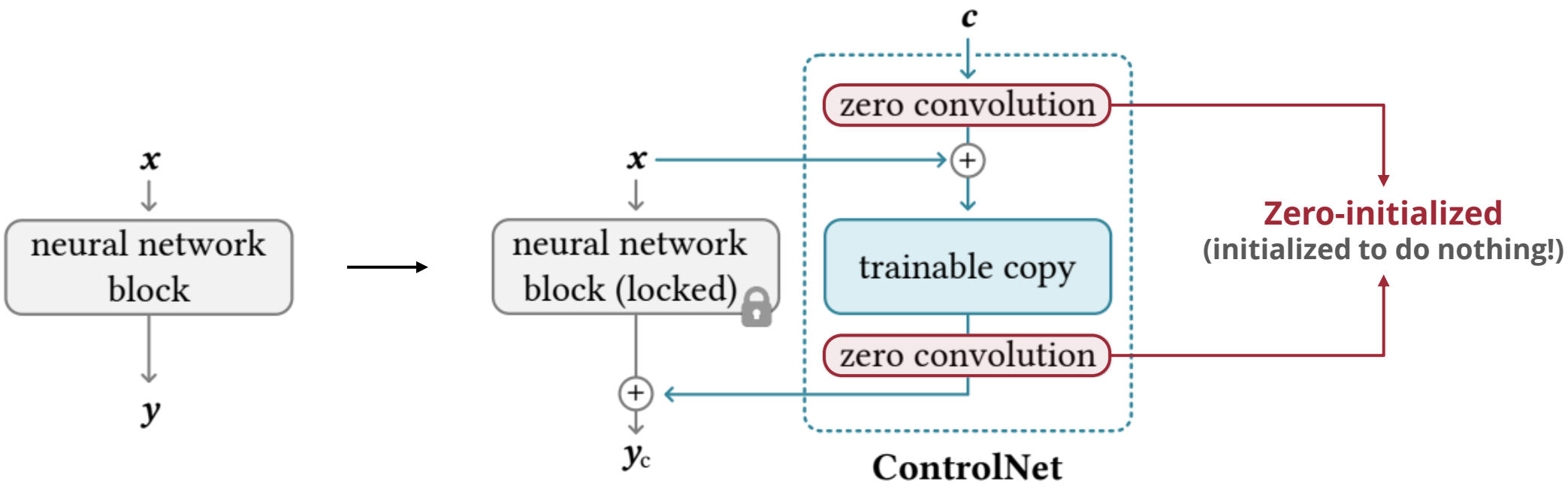


(Source: Zhang et al., 2023)

Can we add controls to a trained text-to-image diffusion model?

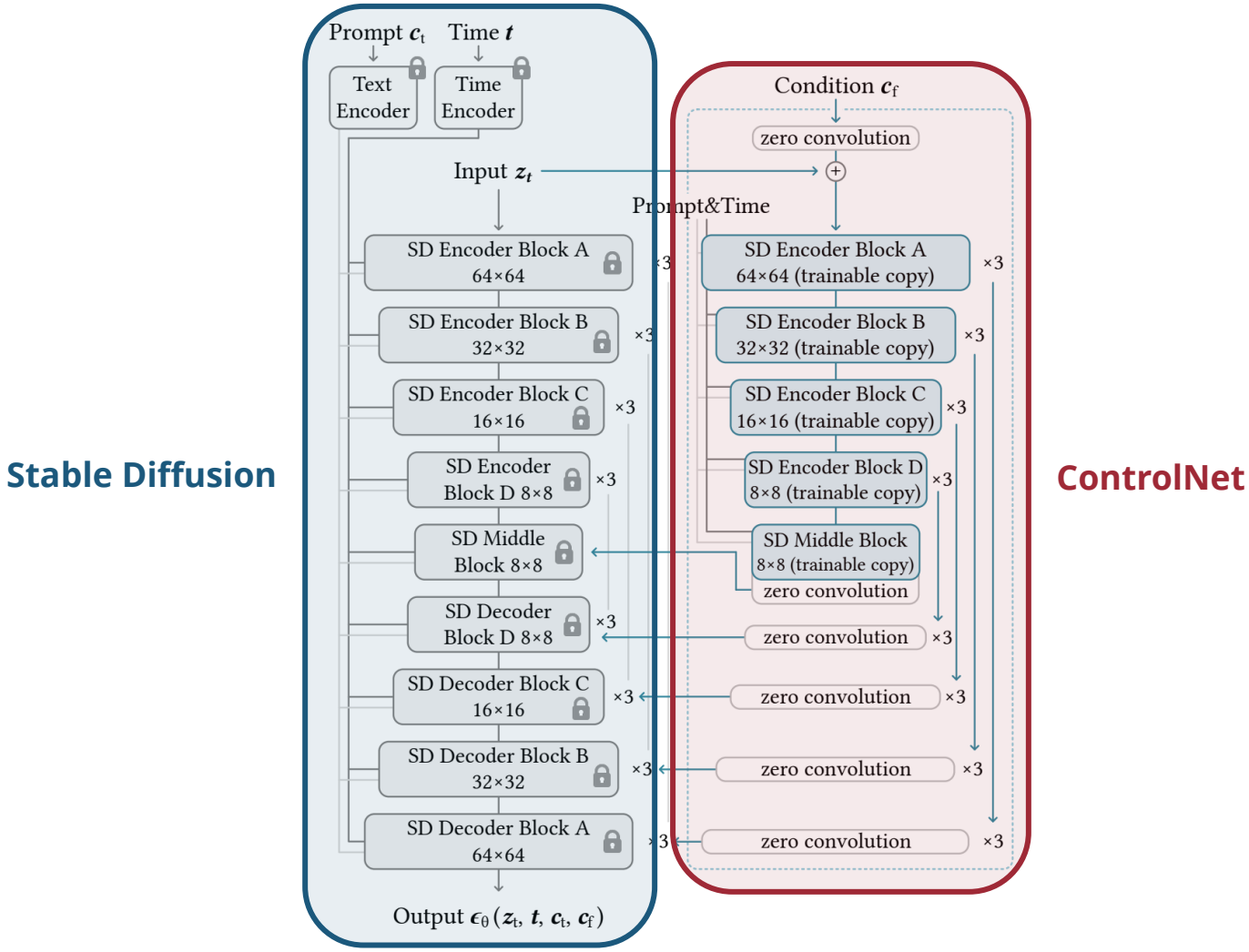
Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," *ICCV*, 2023.

ControlNet (Zhang et al., 2023)



(Source: Zhang et al., 2023)

ControlNet (Zhang et al., 2023)



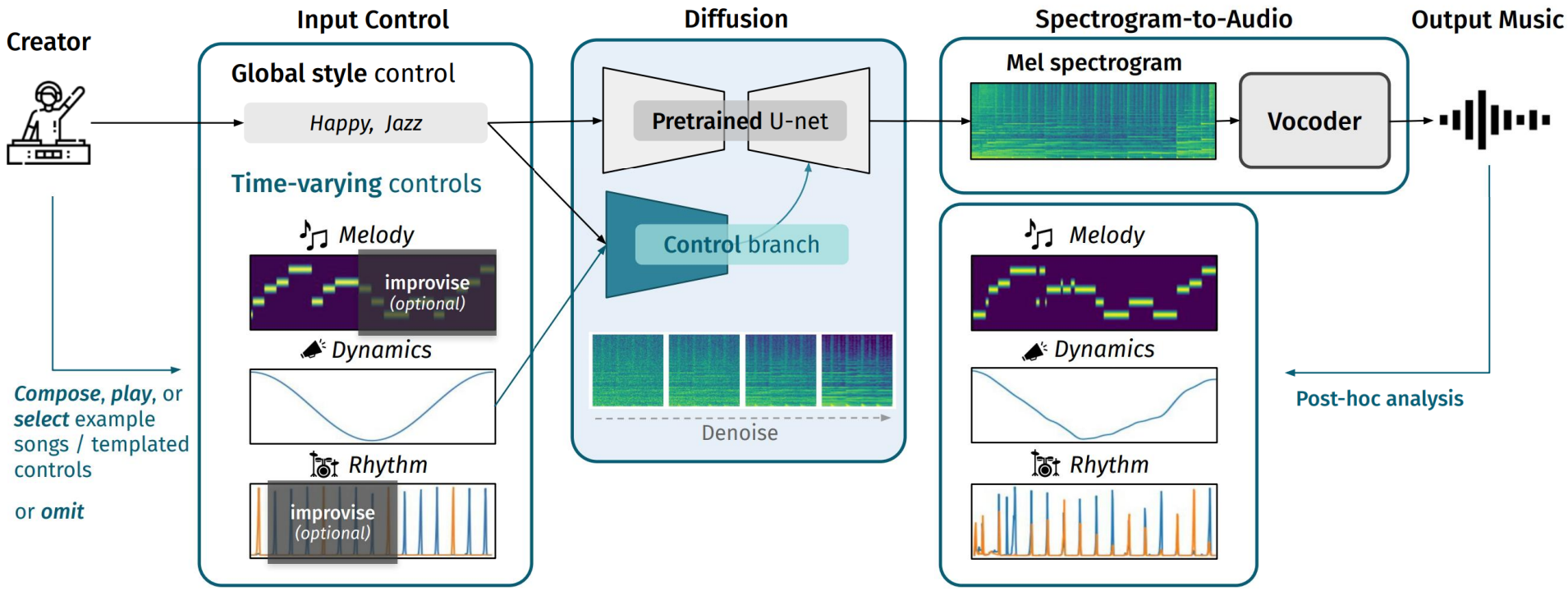
(Source: Zhang et al., 2023)

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," *ICCV*, 2023.

Synthetic Beat Brigade - How would you touch me? (2023)

youtu.be/O4cJ3acEGDw &
drive.google.com/file/d/1QTQ7P3iZI6I0anIwNQ3ewf8g3JjDjesl/view

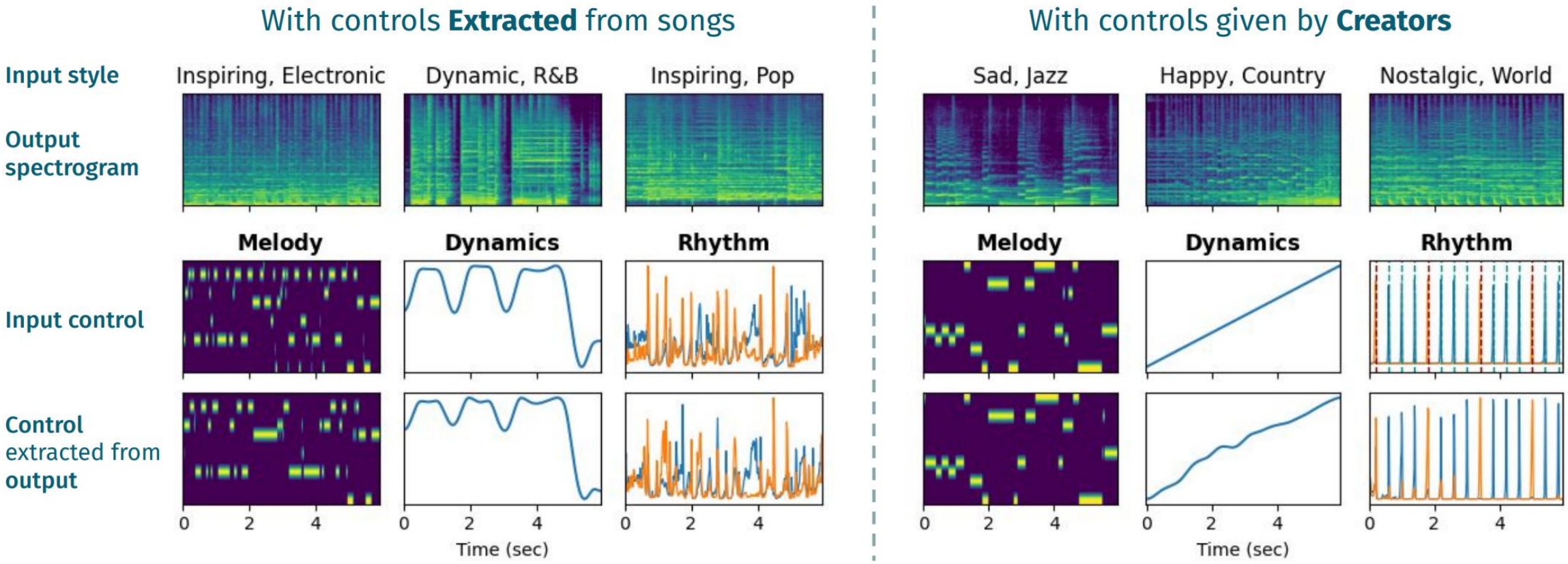
Music ControlNet (Wu et al., 2024)



(Source: Wu et al., 2024)

Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan, "Music ControlNet: Multiple Time-varying Controls for Music Generation," *TASLP*, 2024.

Music ControlNet (Wu et al., 2024)

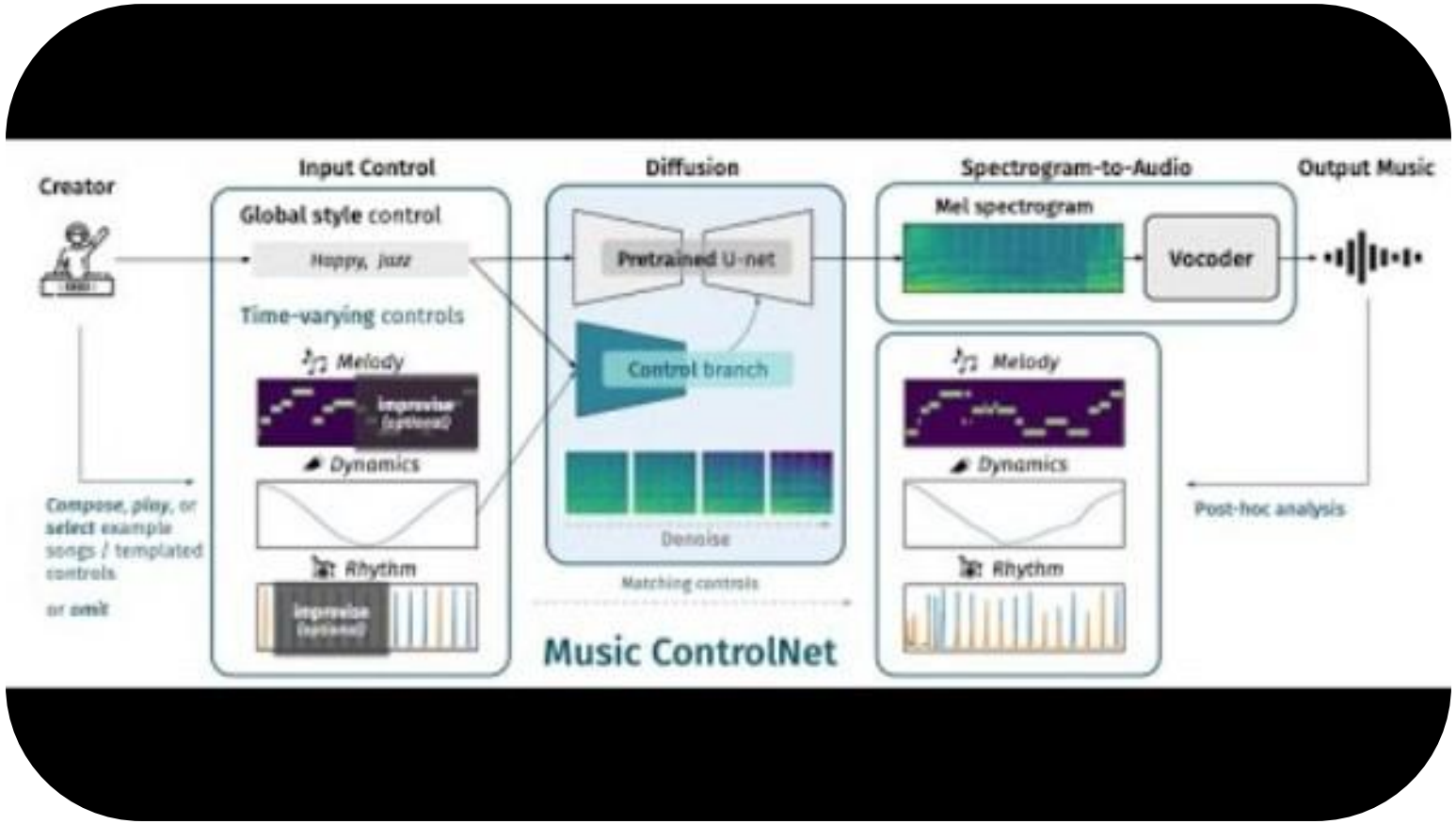


(Source: Wu et al., 2024)

musiccontrolnet.github.io/web

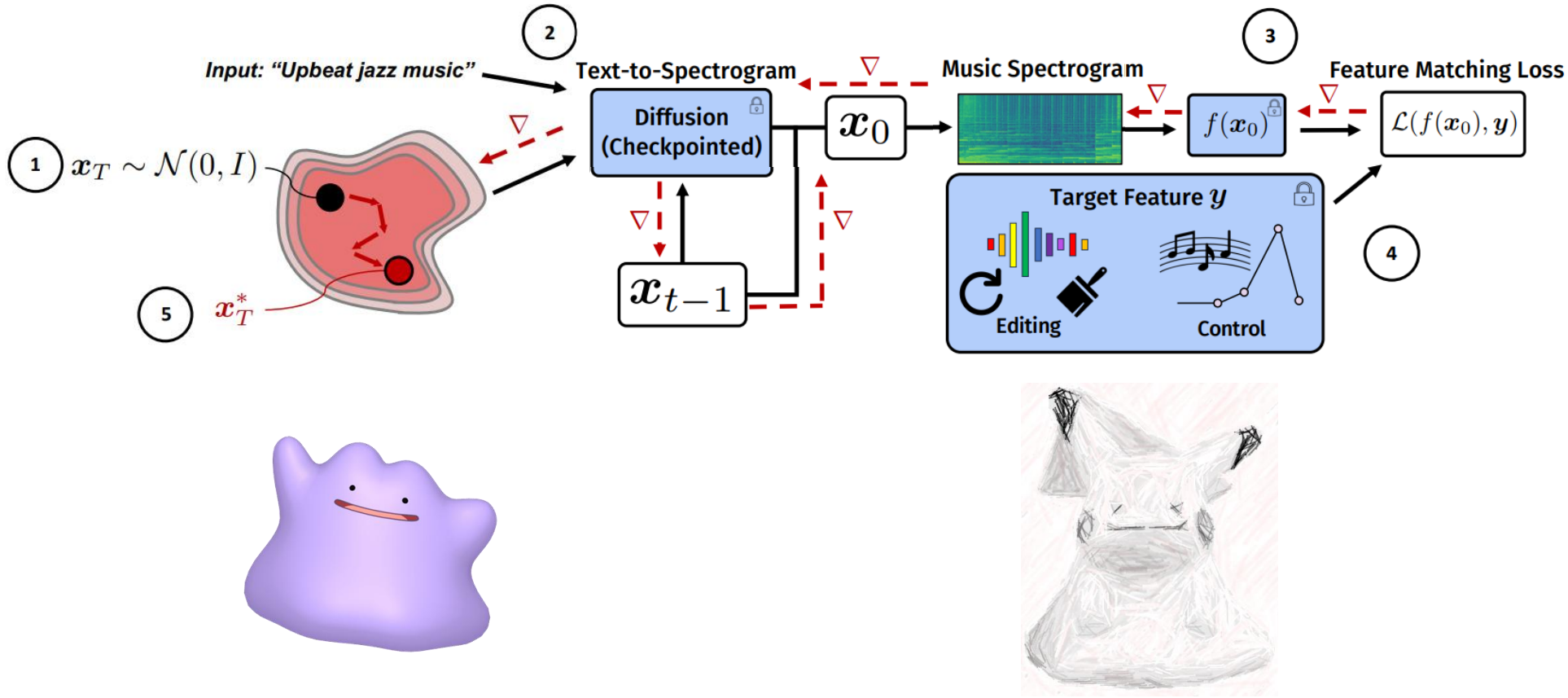
Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan, "Music ControlNet: Multiple Time-varying Controls for Music Generation," *TASLP*, 2024.

Music ControlNet (Wu et al., 2024)



youtu.be/QVr-S-DyccU

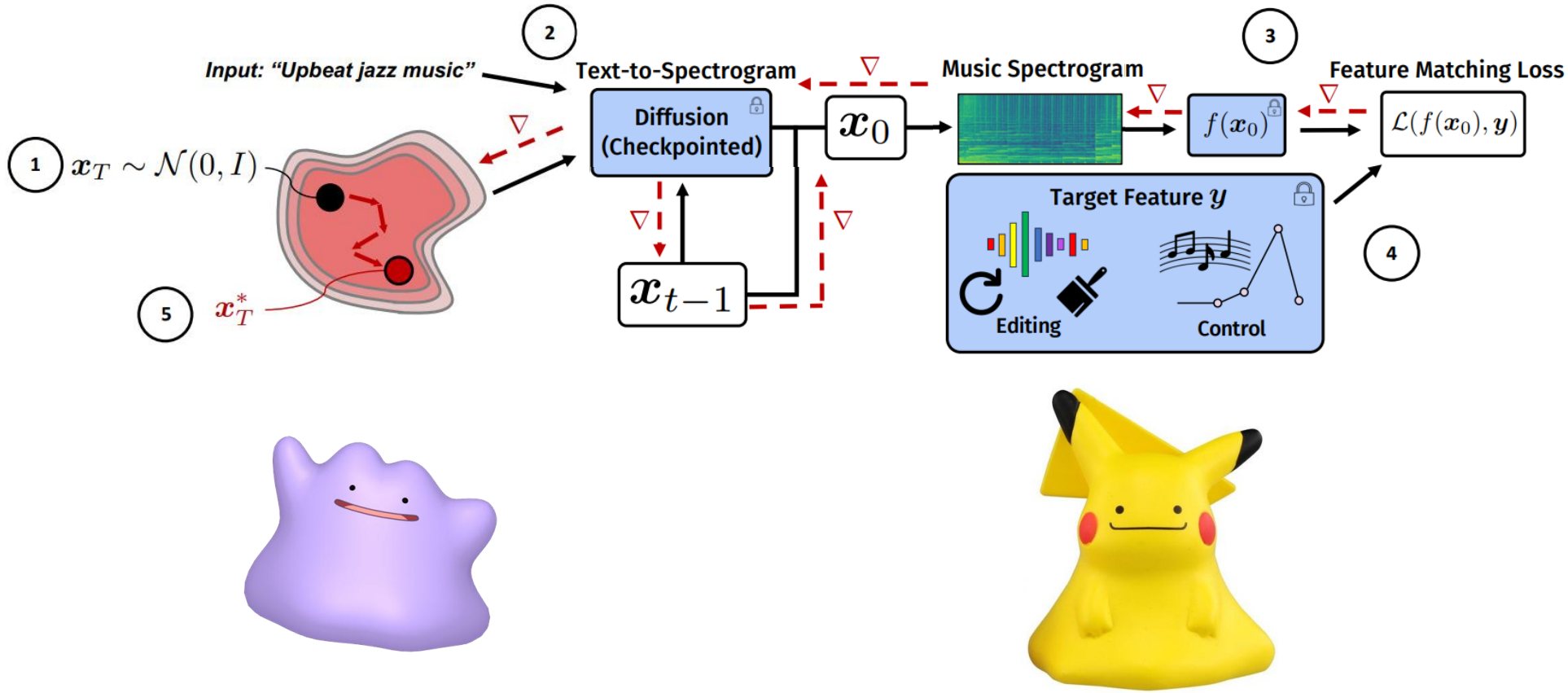
DITTO (Novack et al., 2024)



(Source: Novack et al., 2024)

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan, "DITTO: Diffusion Inference-Time T-Optimization for Music Generation," *ICML*, 2024.

DITTO (Novack et al., 2024)

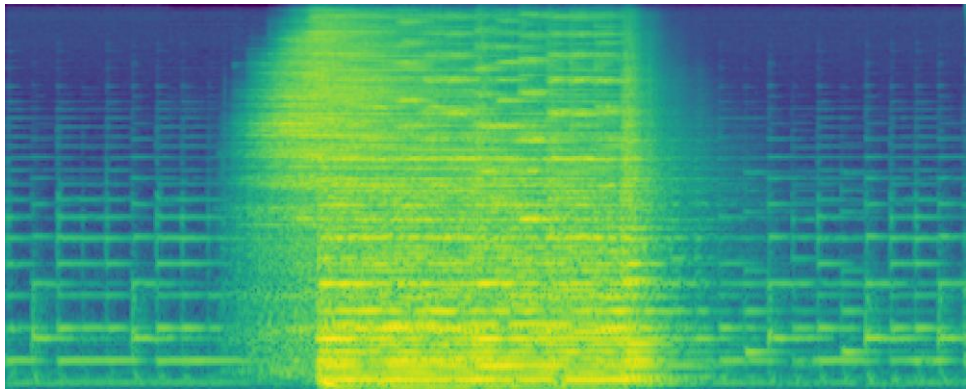


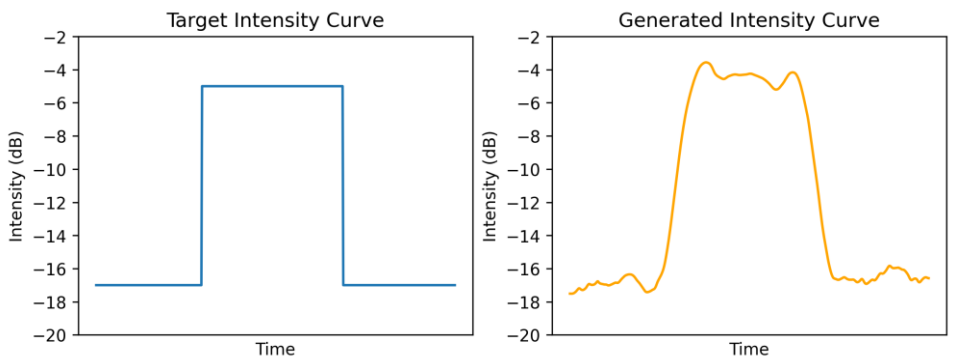
(Source: Novack et al., 2024)

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan, "DITTO: Diffusion Inference-Time T-Optimization for Music Generation," *ICML*, 2024.

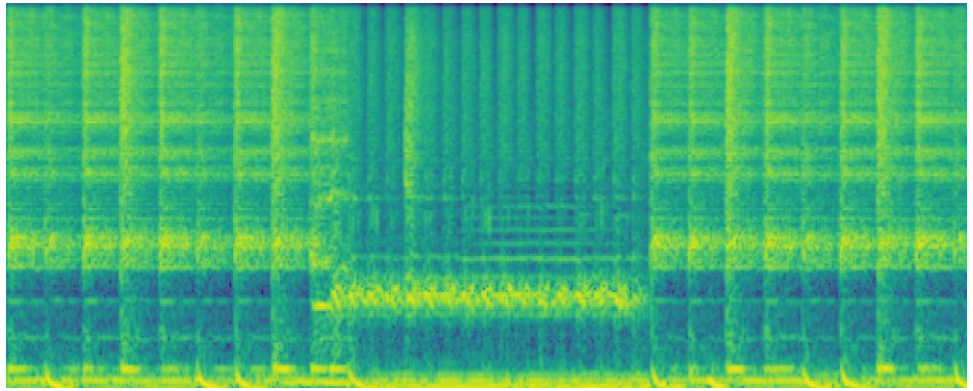
DITTO (Novack et al., 2024)

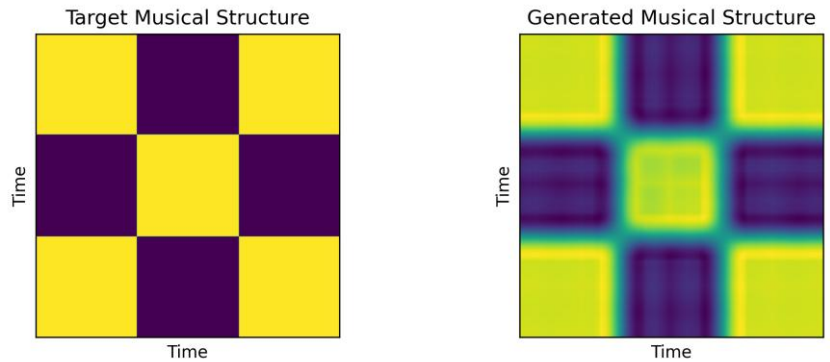
Intensity control





Structure control





(Source: Novack et al., 2024)

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan, "DITTO: Diffusion Inference-Time T-Optimization for Music Generation," *ICML*, 2024.

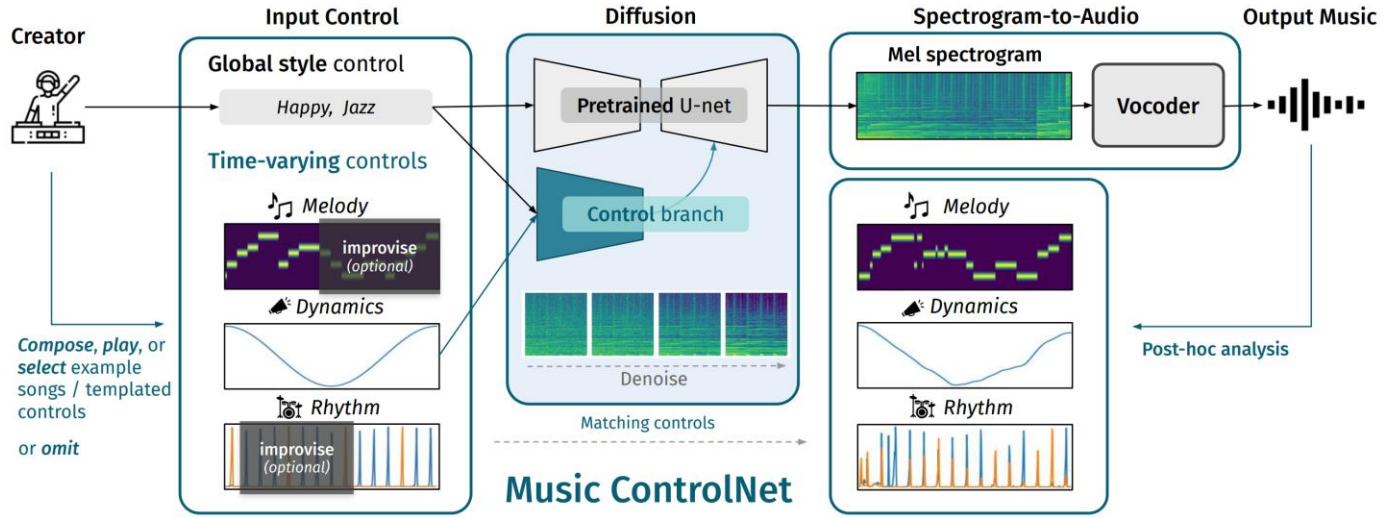
DITTO (Novack et al., 2024)

youtu.be/KooosSNPNo8 & ditto-music.github.io/web/

Music ControlNet vs DITTO

Music ControlNet

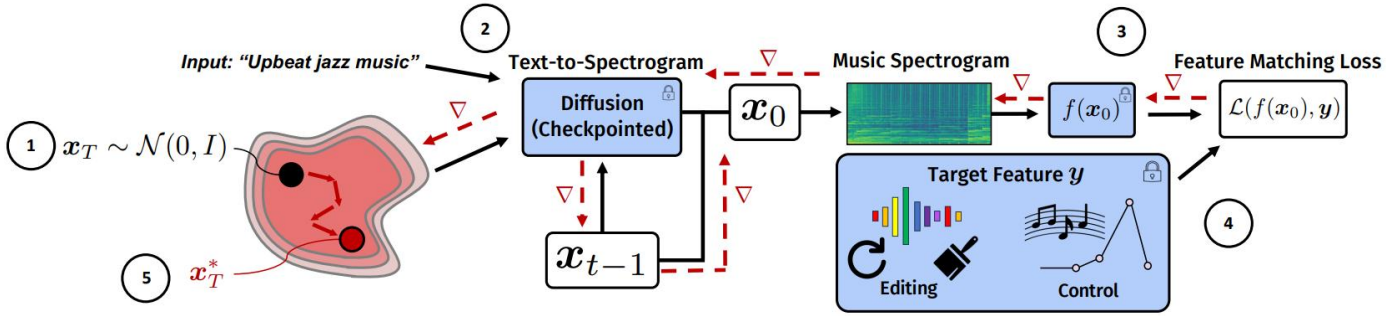
Needs some training!



(Source: Wu et al., 2024)

DITTO

No training needed!



(Source: Novack et al., 2024)

Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan, "Music ControlNet: Multiple Time-varying Controls for Music Generation," *TASLP*, 2024.
Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan, "DITTO: Diffusion Inference-Time T-Optimization for Music Generation," *ICML*, 2024.