PAT 464/564 (Winter 2026)

Generative AI for Music & Audio Creation

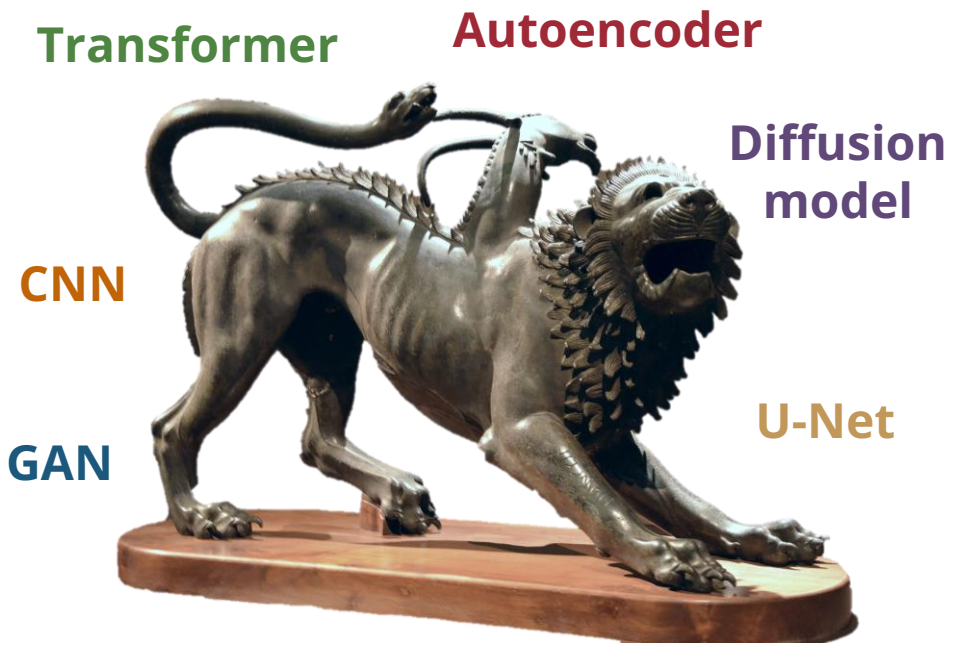
Lecture 15: Latent Diffusion Models

Instructor: Hao-Wen Dong

UNIVERSITY OF MICHIGAN

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

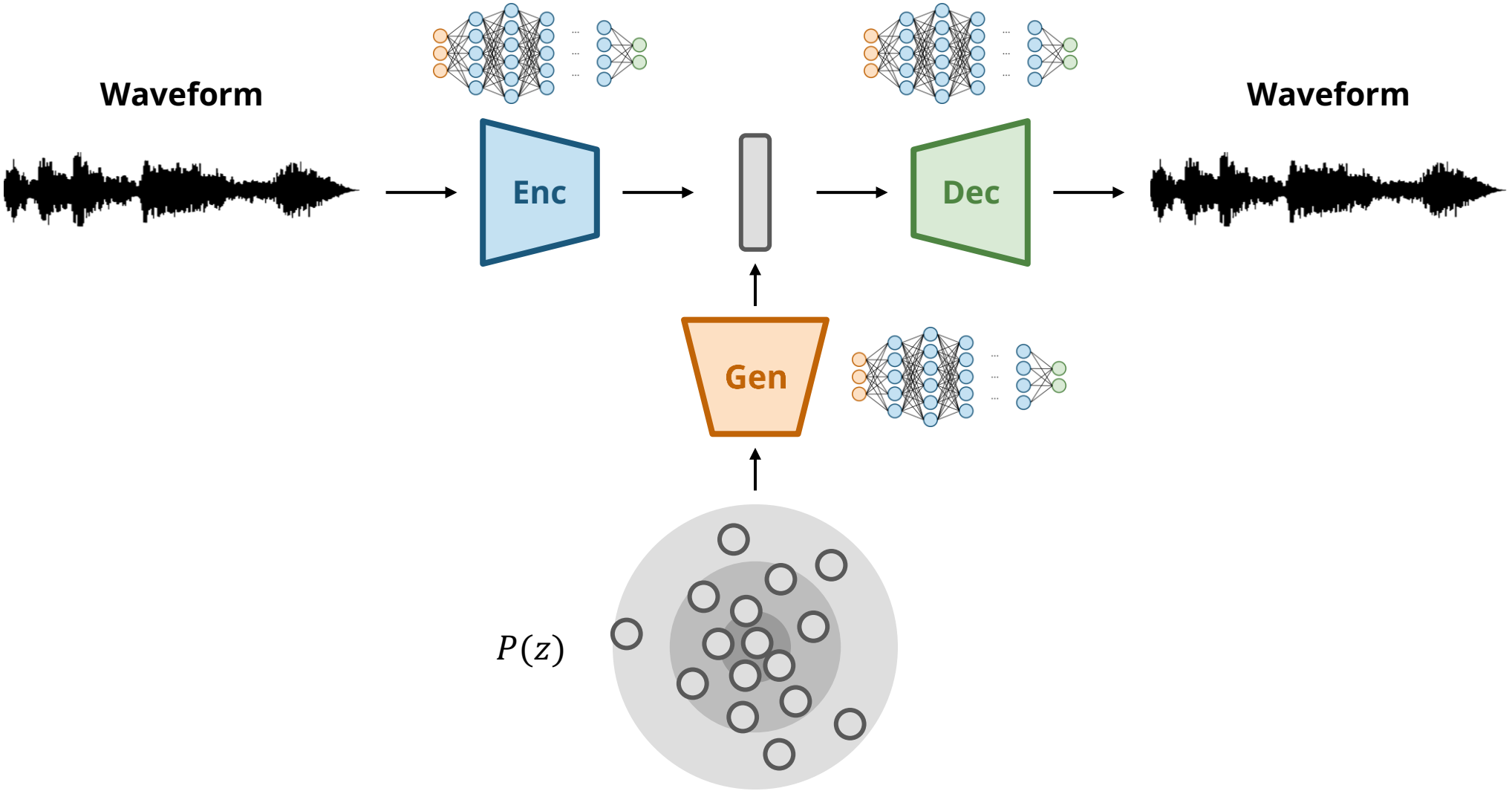


(Source: Raddato via worldhistory.org)

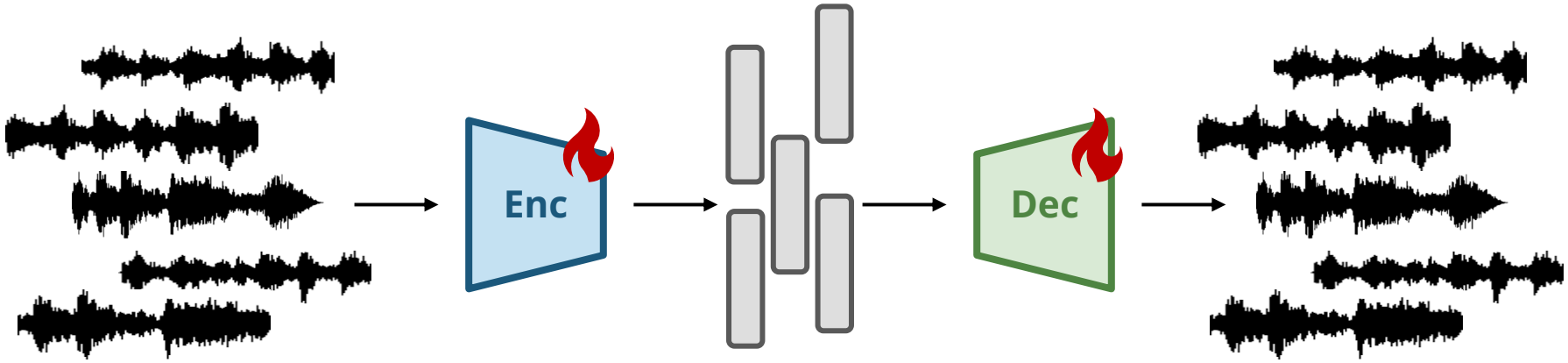
worldhistory.org/image/5462/chimera-of-arezzo-florence/

Latent-based Music Synthesis

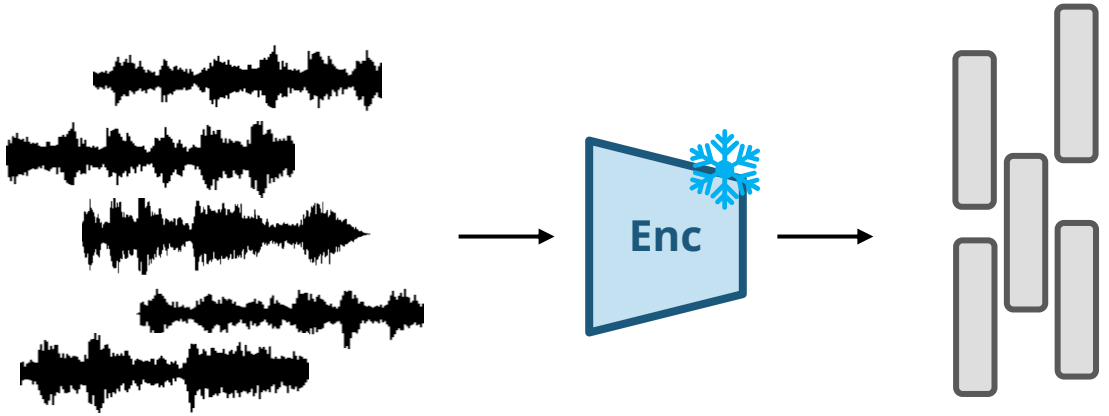
Latent-based Audio Synthesis



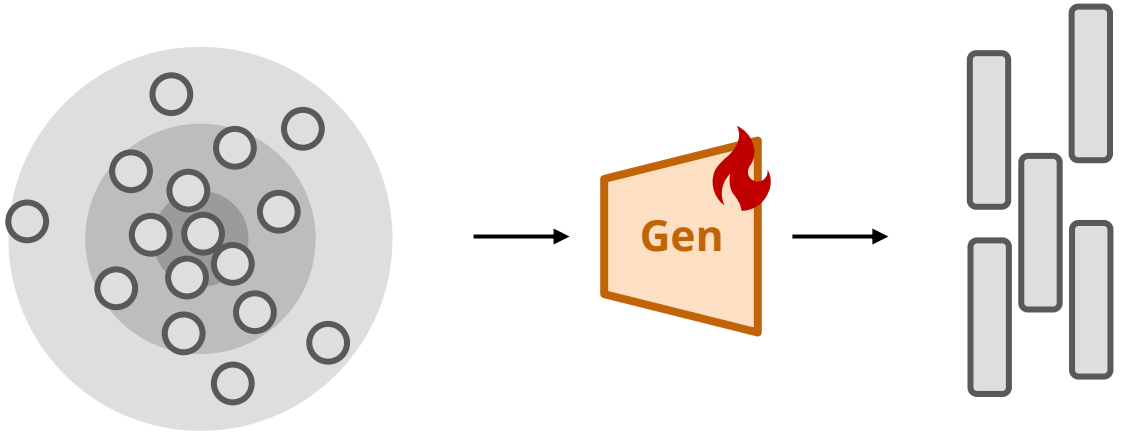
Step 1: Train an Autoencoder



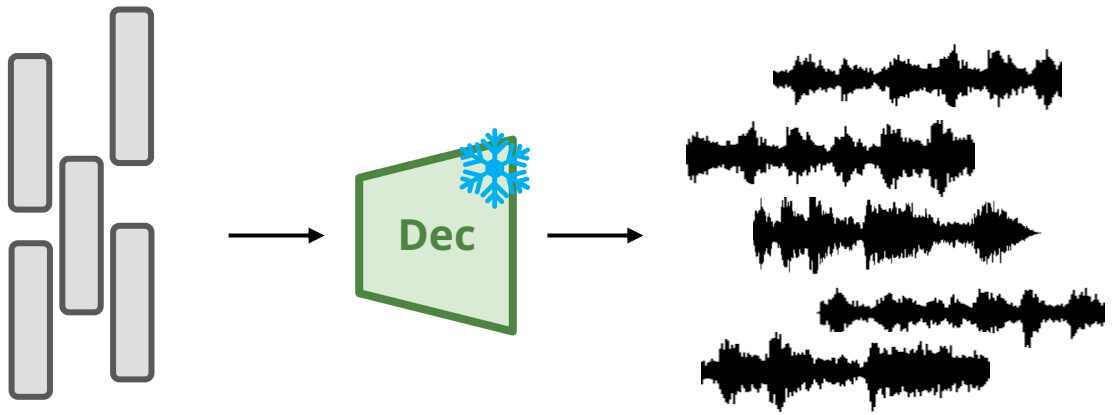
Step 2: Compute the Latent Vectors



Step 3: Train a Latent Generative Model

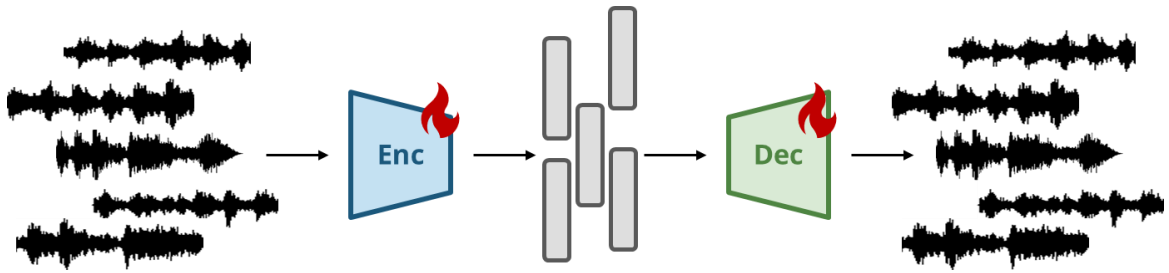


Step 4: Decode the Latent Vectors

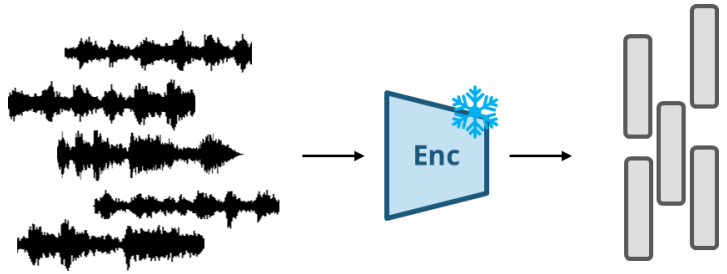


Latent-based Audio Synthesis: Pipeline

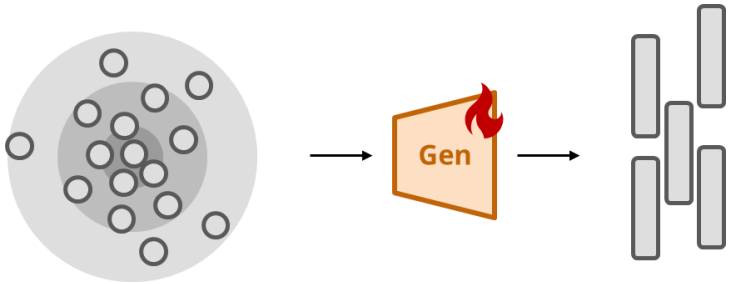
Step 1: Train an Autoencoder



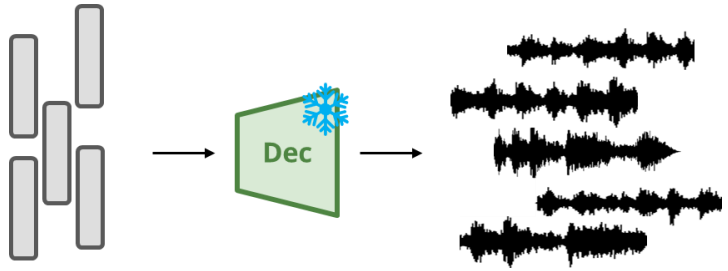
Step 2: Compute the Latent Vectors



Step 3: Train a Latent Generative Model

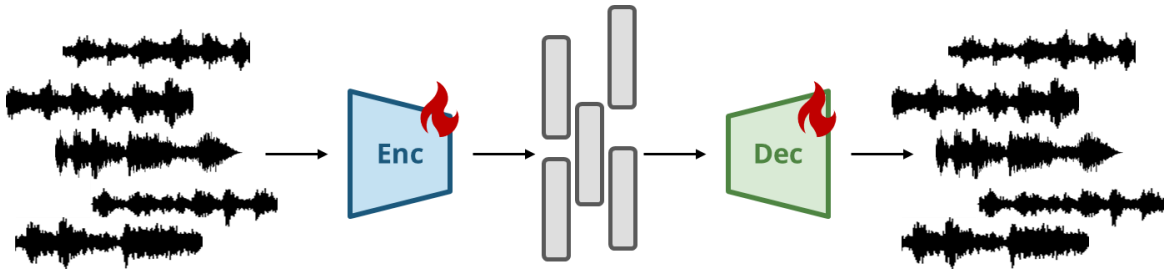


Step 4: Decode the Latent Vectors

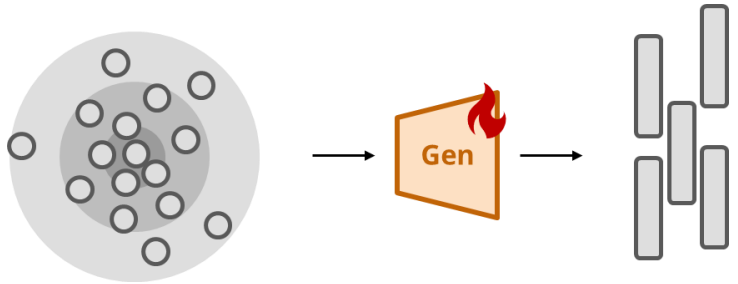


Latent-based Audio Synthesis: Training

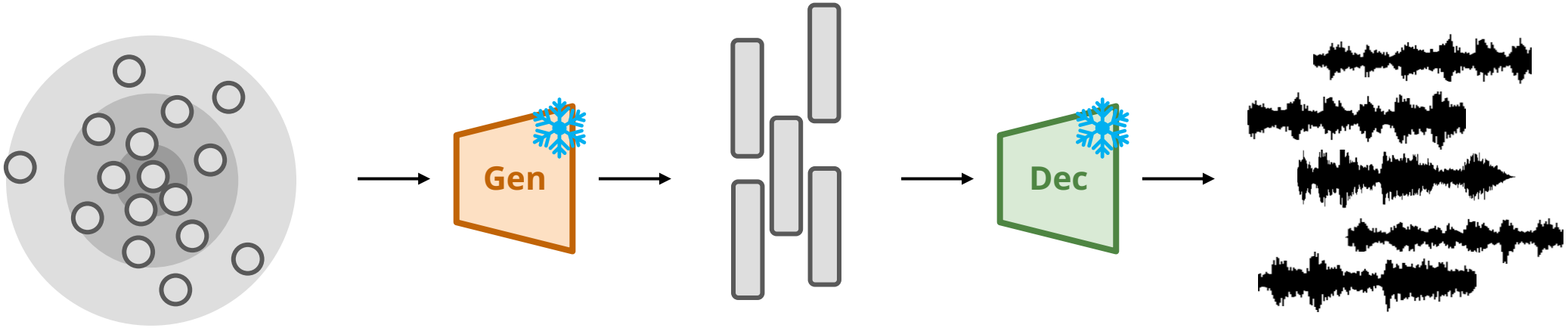
Autoencoder



Latent Generative Model

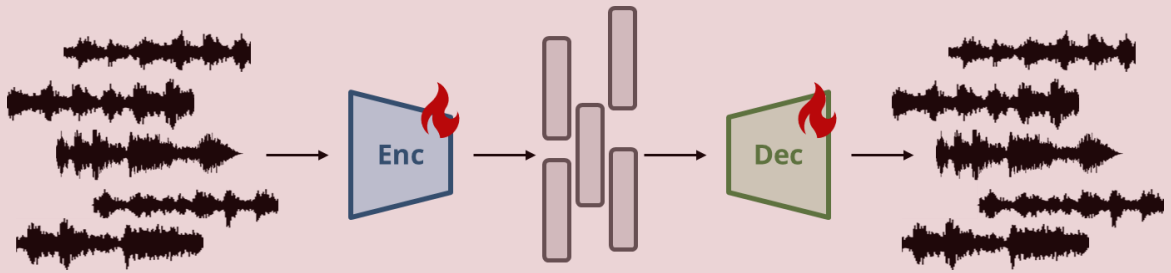


Latent-based Audio Synthesis: Inference

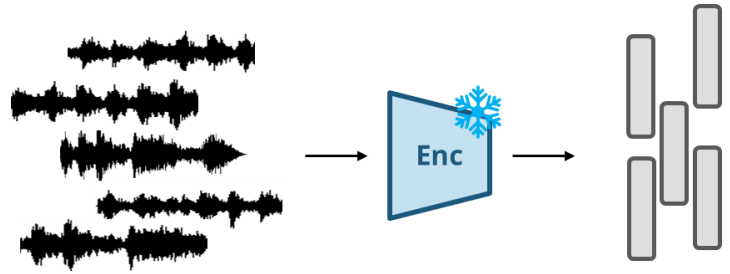


Latent-based Audio Synthesis: Pipeline

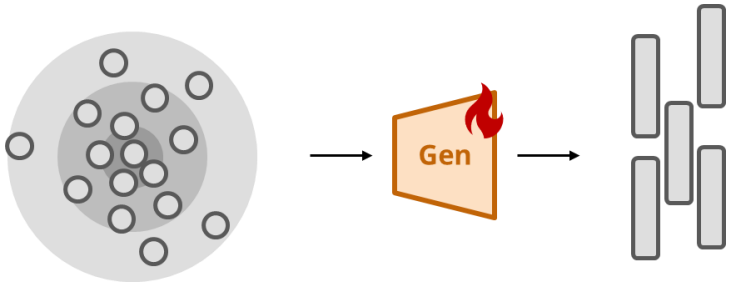
Step 1: Train an Autoencoder



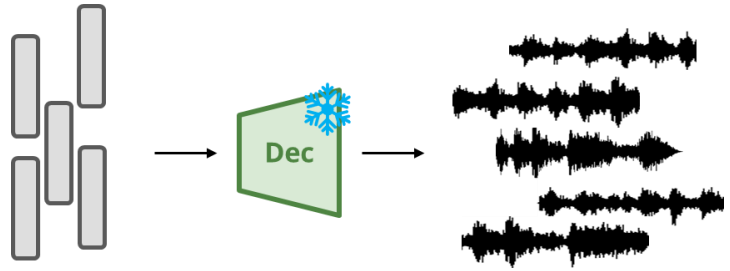
Step 2: Compute the Latent Vectors



Step 3: Train a Latent Generative Model

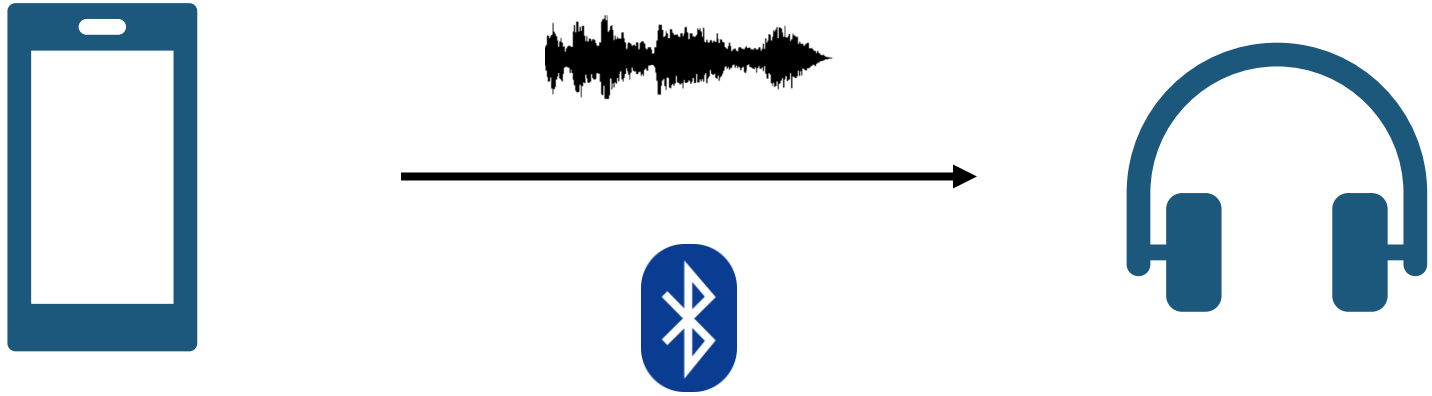


Step 4: Decode the Latent Vectors



Neural Codecs

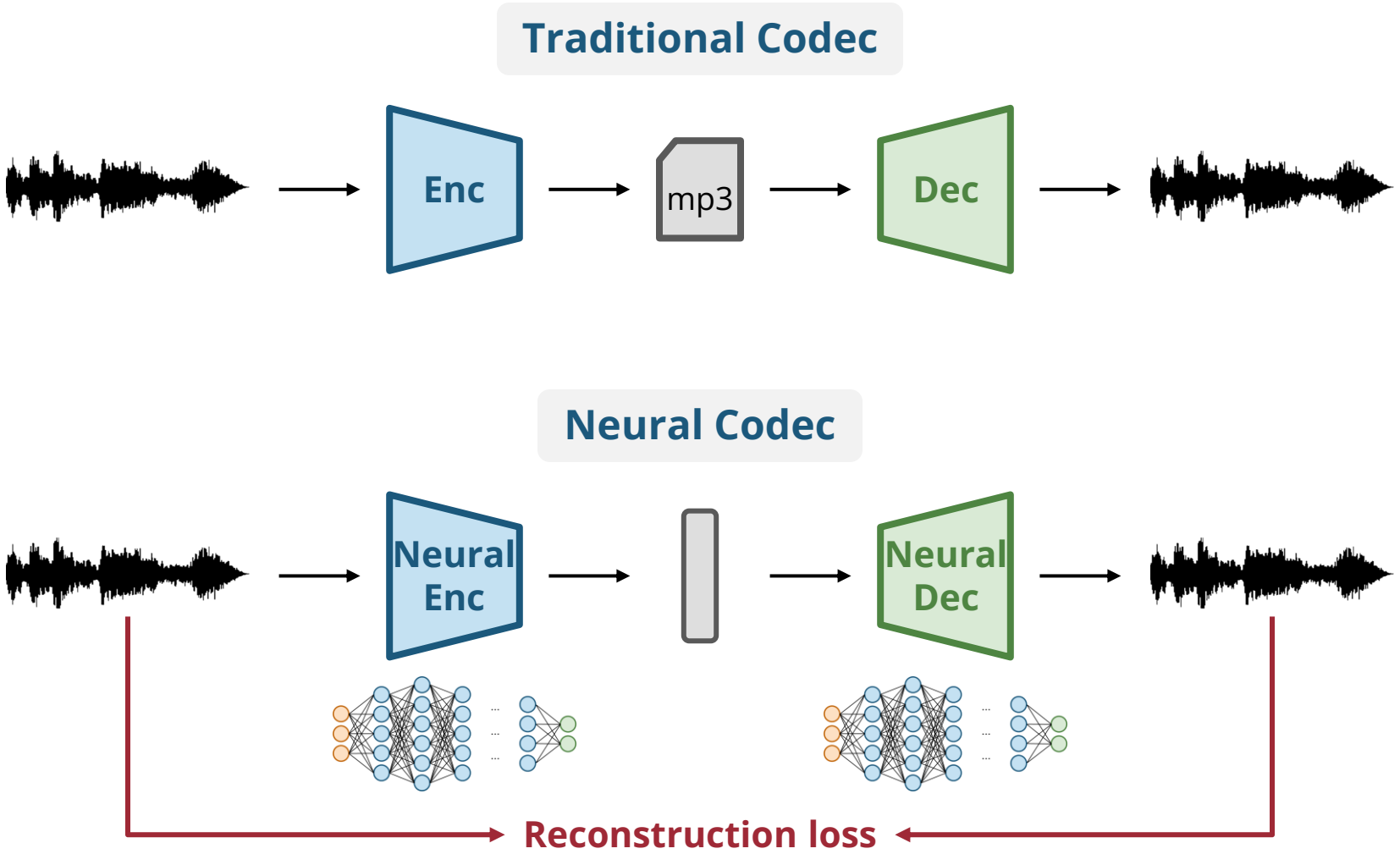
What is a Codec?



SBC, AAC, aptX,
aptX HD, LDAC

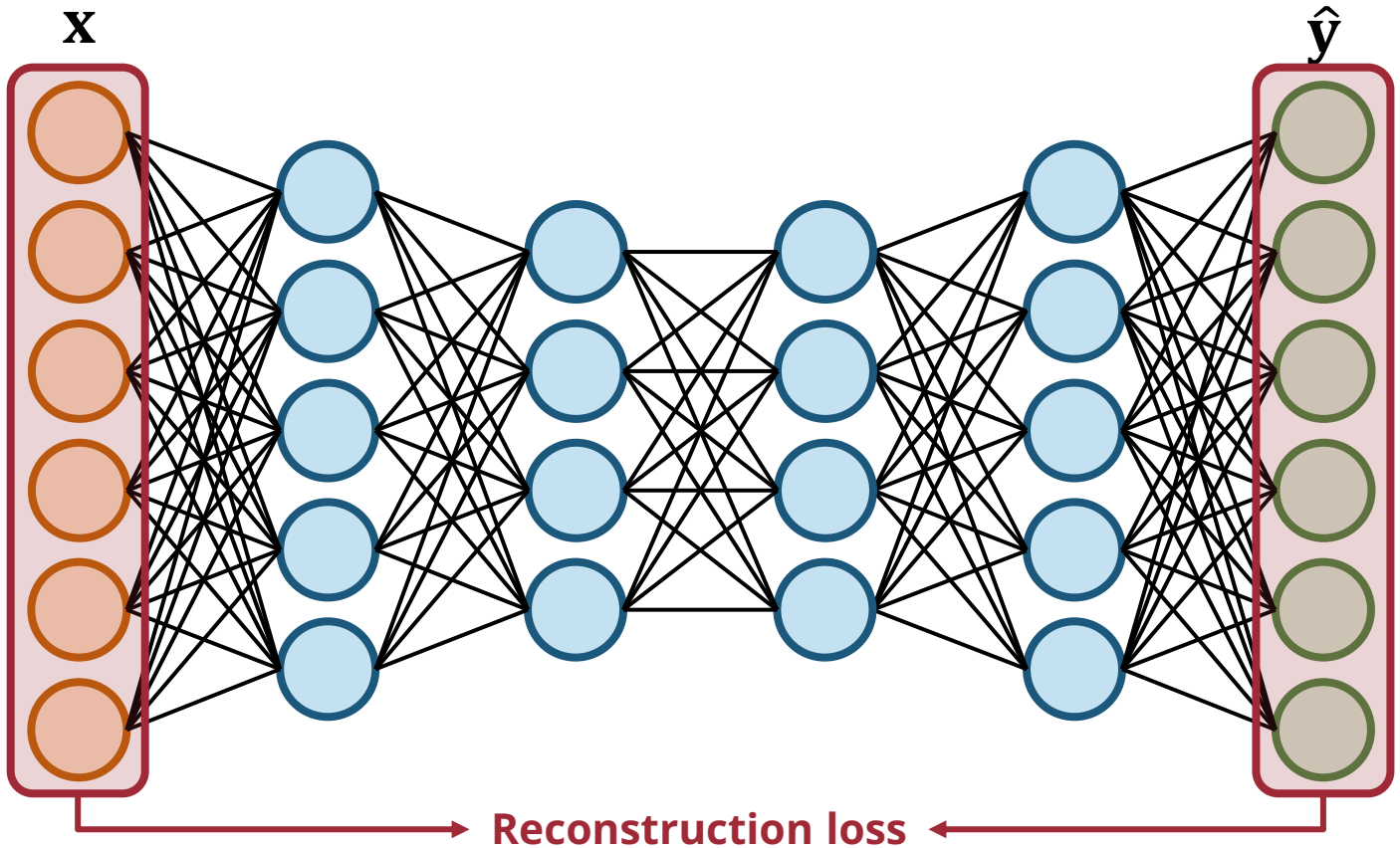
What is a Codec?

Neural Codec

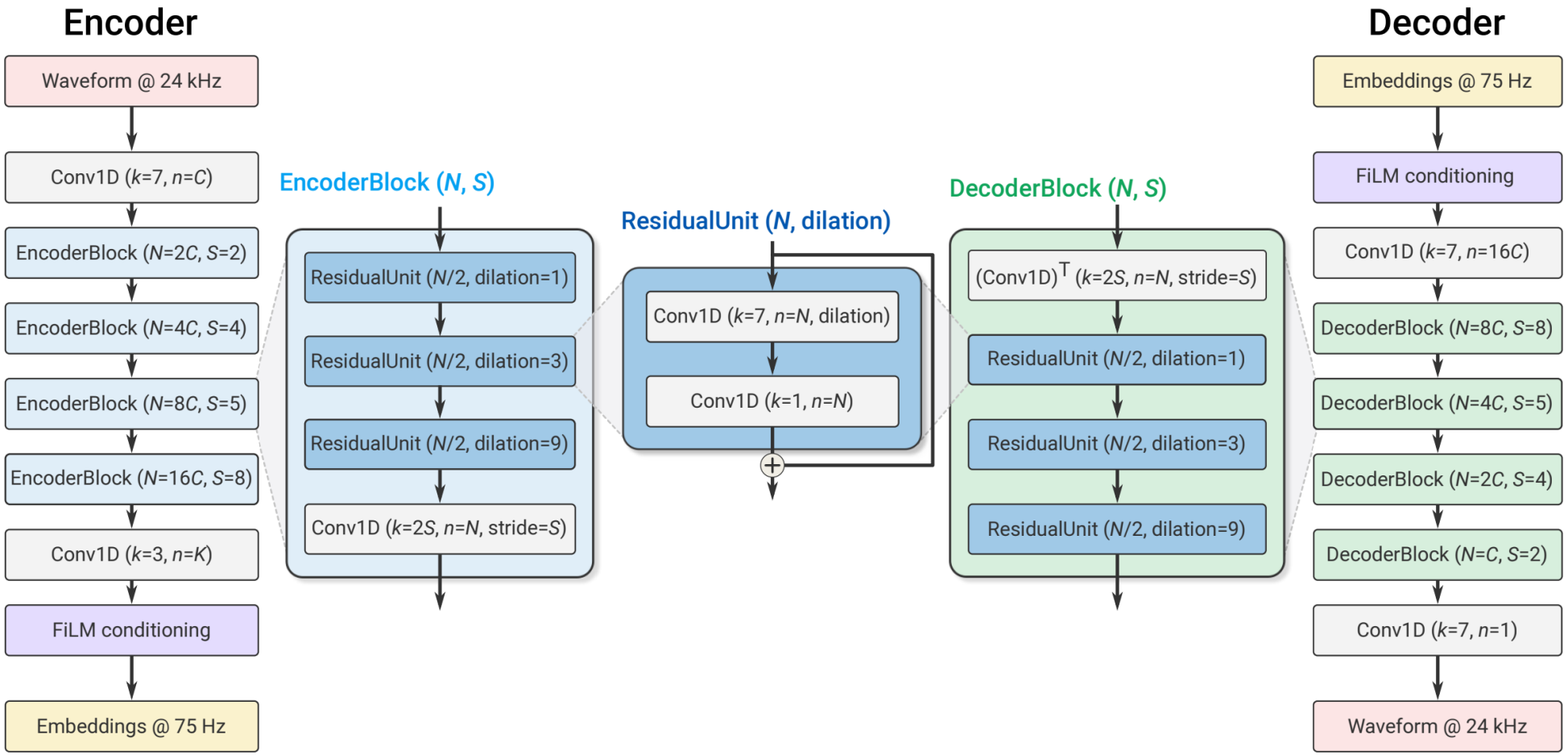


Autoencoders

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



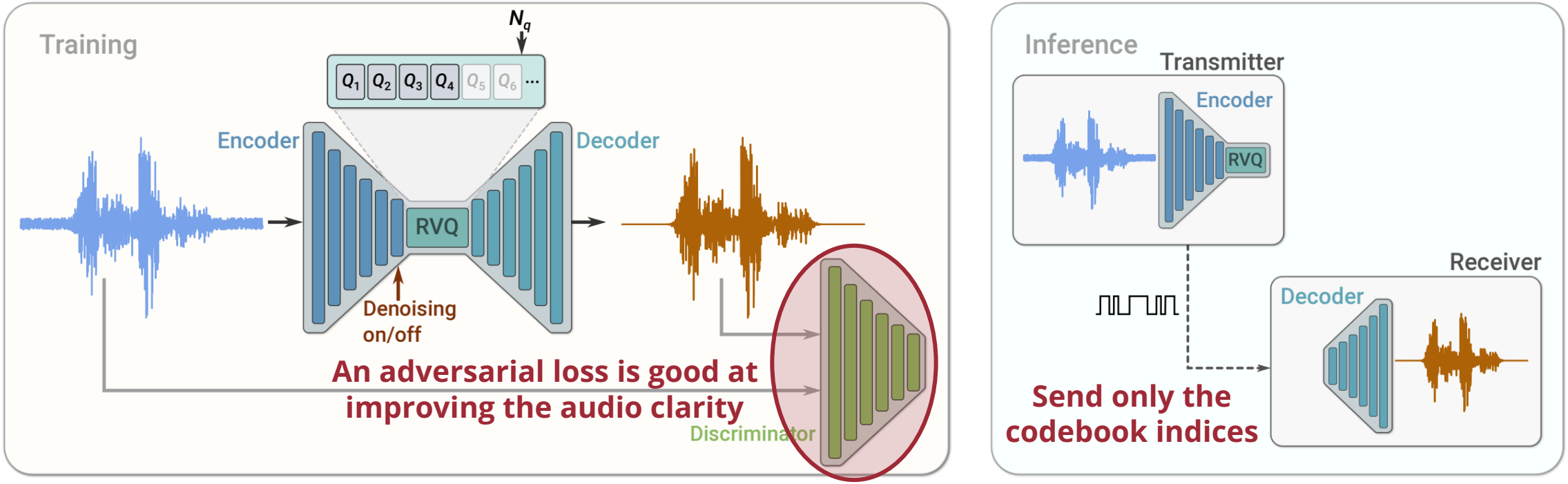
SoundStream (Zeghidour et al., 2021)



(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.

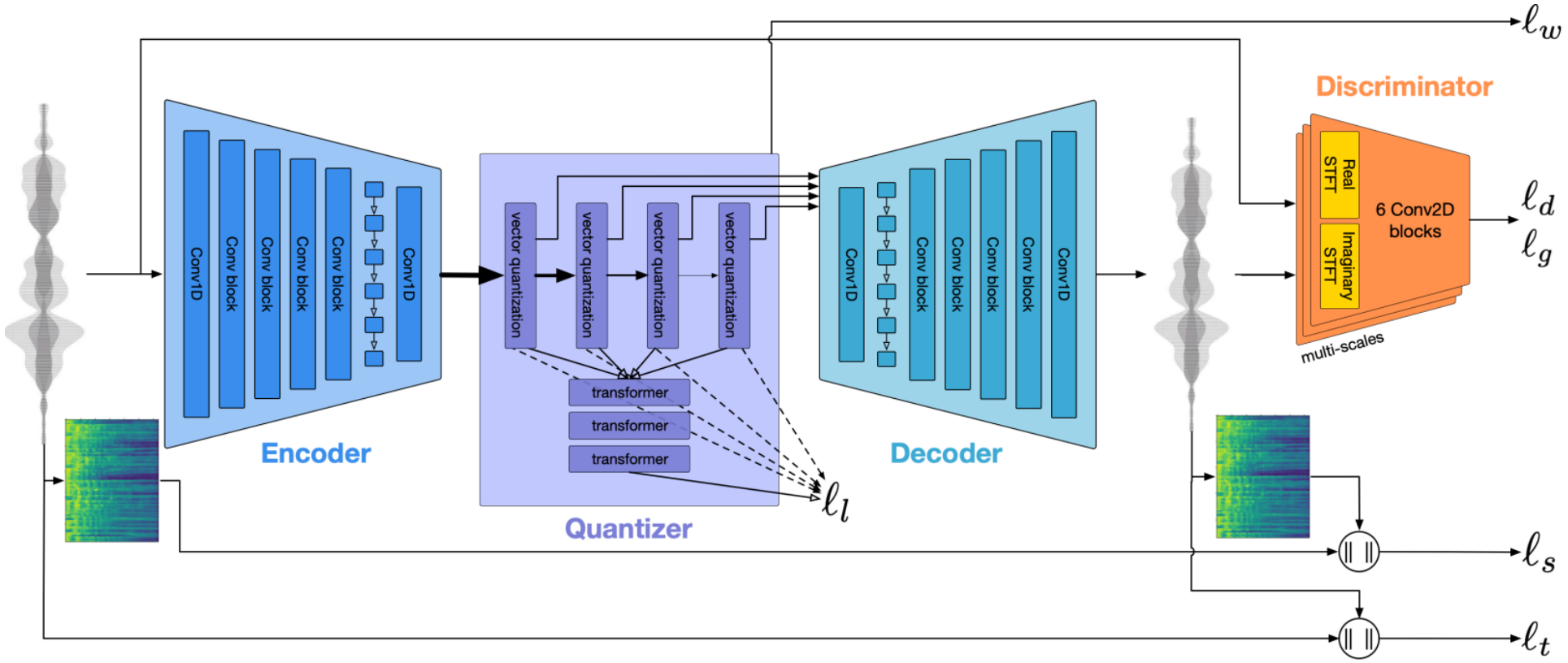
SoundStream (Zeghidour et al., 2021)



(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)



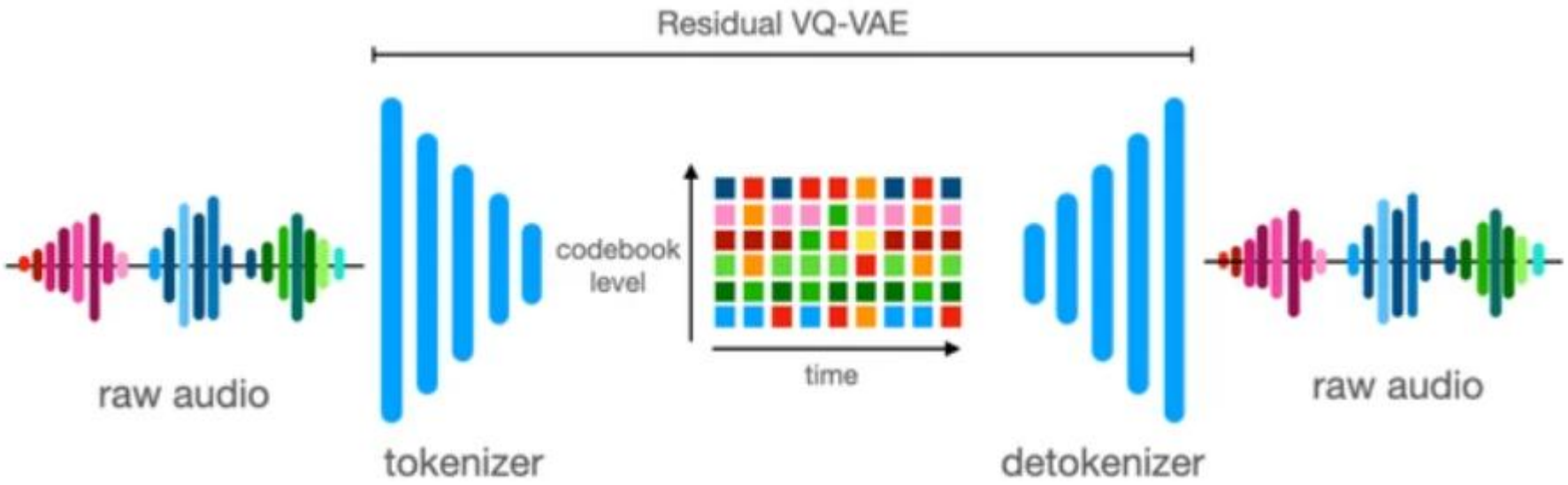
(Source: Défossez et al., 2022)

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)

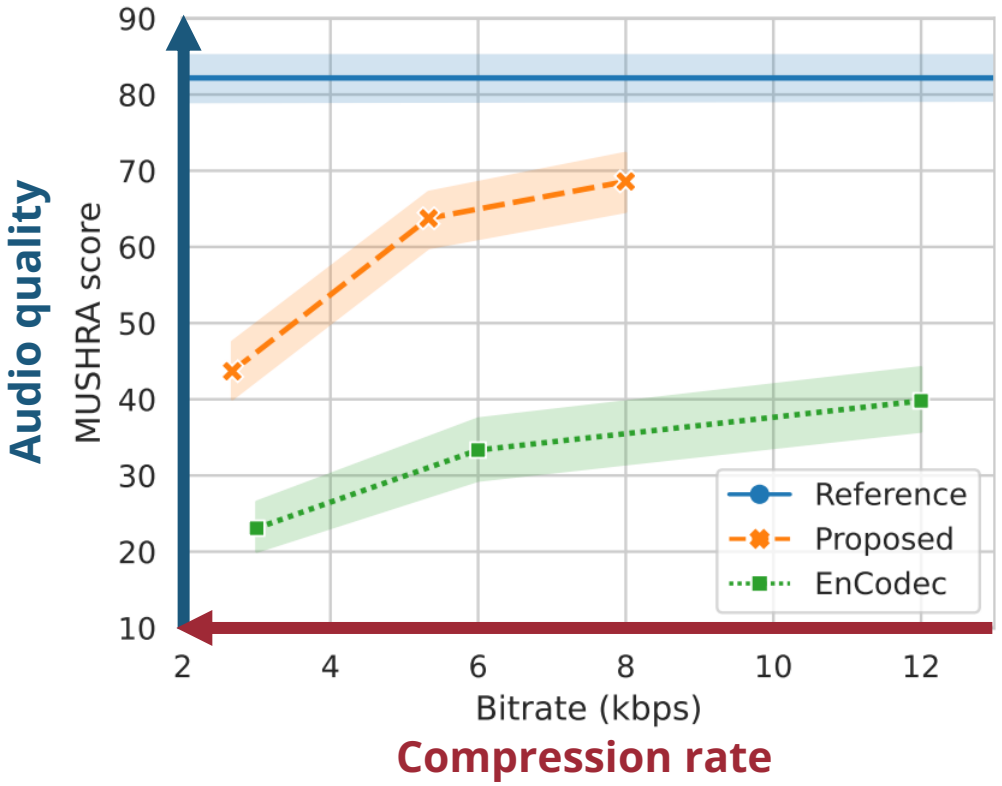
descript.notion.site/Descript-Audio-Codec-11389fce0ce2419891d6591a68f814d5

github.com/descriptinc/descript-audio-codec

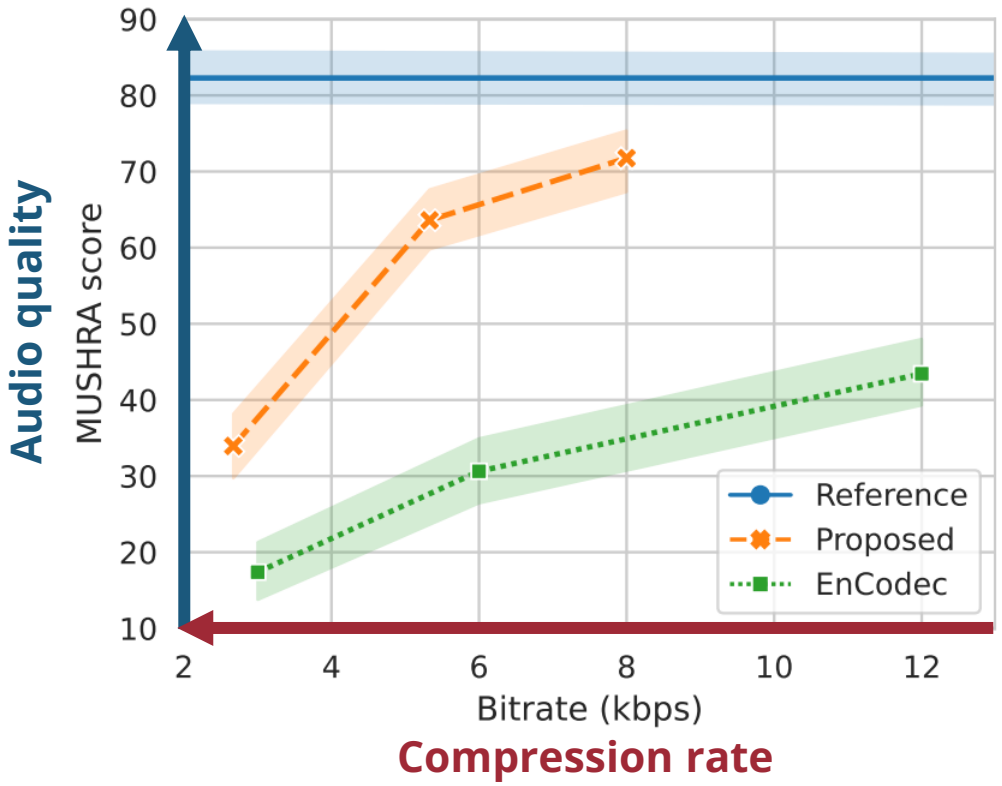
Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-Fidelity Audio Compression with Improved RVQGAN," *NeurIPS*, 2023.

Descript Audio Codec (Kumar et al., 2023)

Listening Test Results @ 44.1 kHz



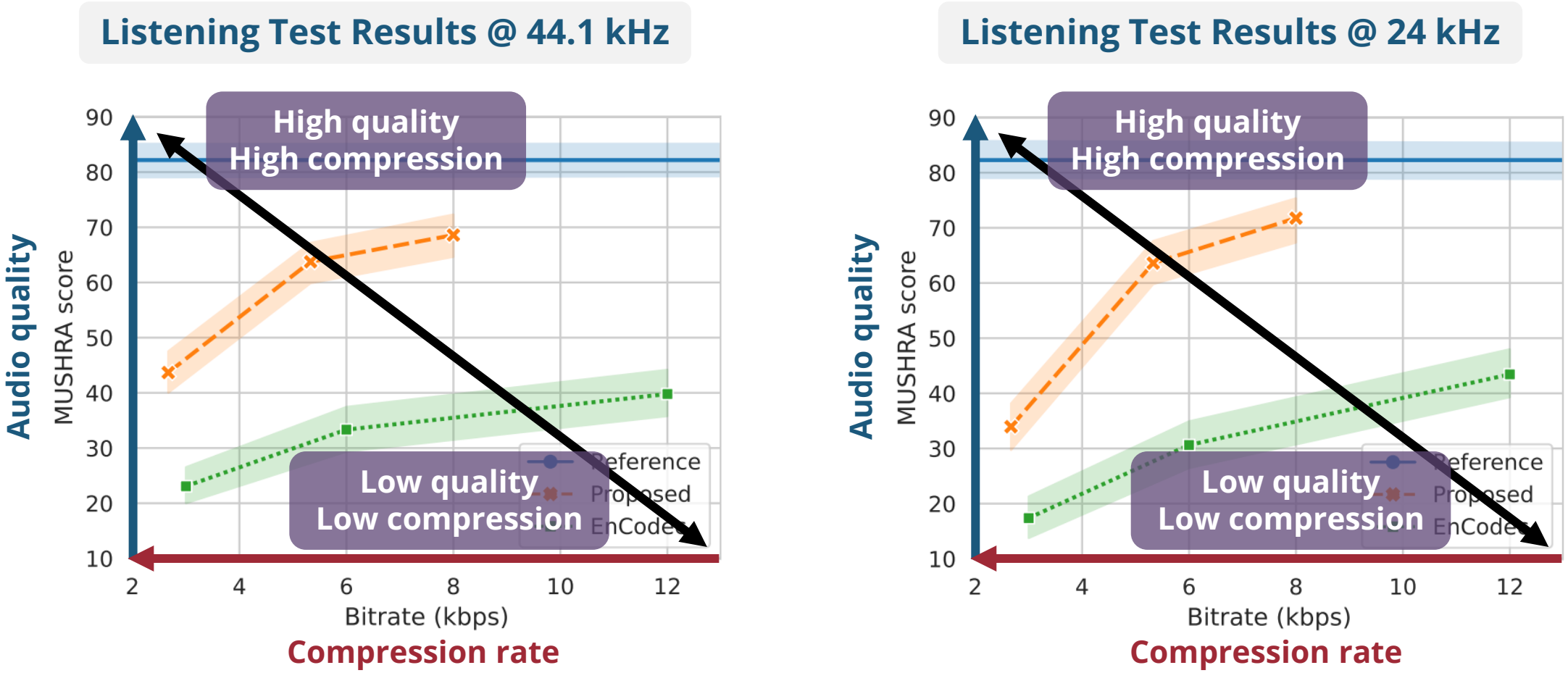
Listening Test Results @ 24 kHz



(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)



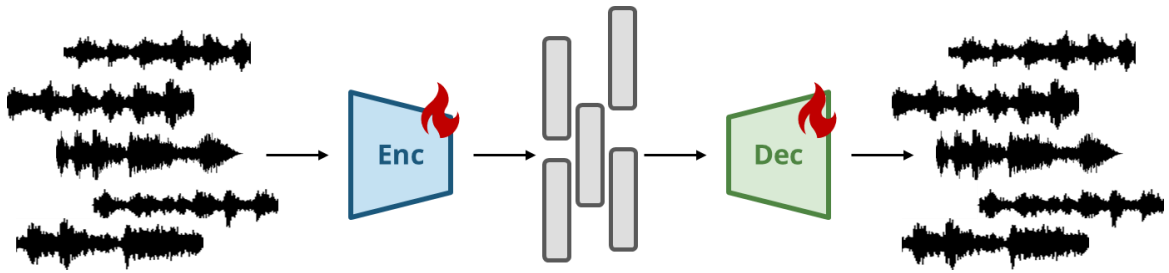
(Source: Kumar et al., 2023)

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

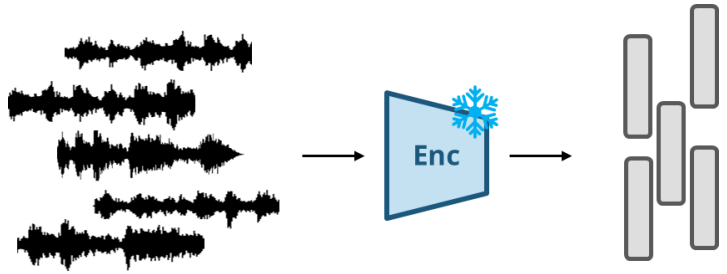
Latent-based Text-to-Music Synthesis

Latent-based Audio Synthesis: Pipeline

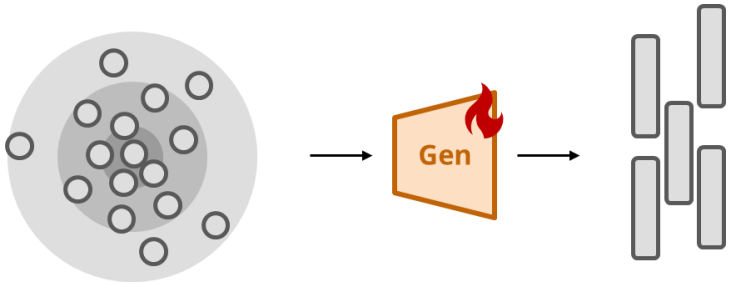
Step 1: Train an Autoencoder



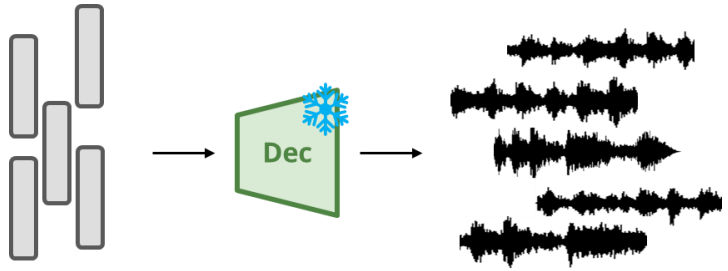
Step 2: Compute the Latent Vectors



Step 3: Train a Latent Generative Model

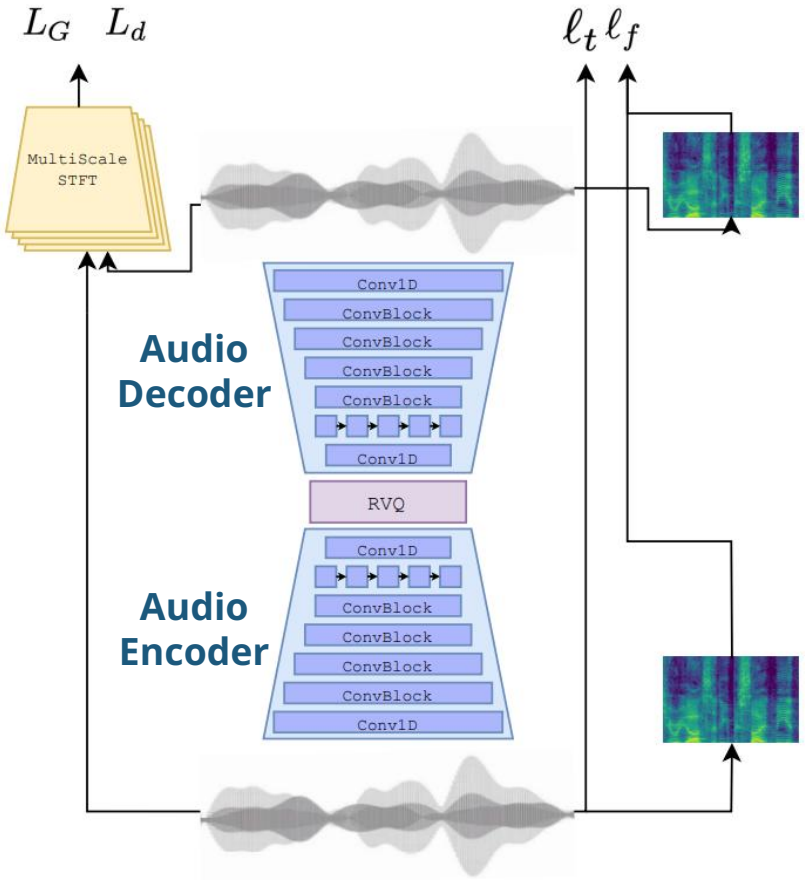


Step 4: Decode the Latent Vectors

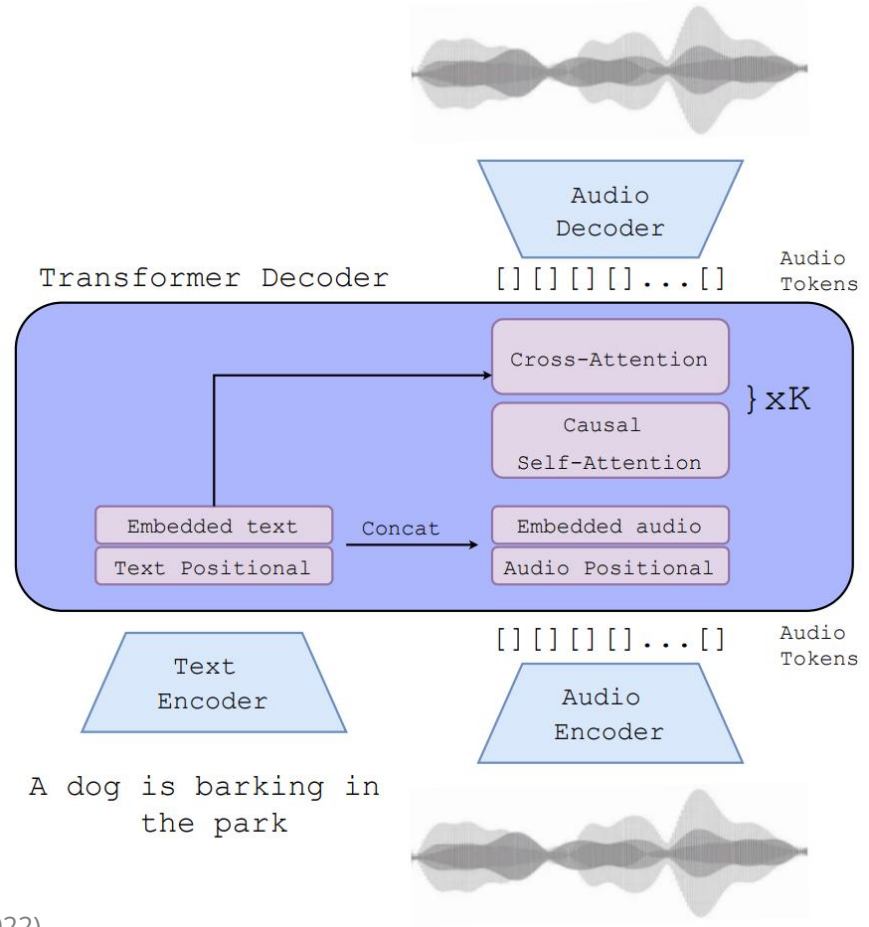


AudioGen (Kreuk et al., 2023)

Audio Autoencoder



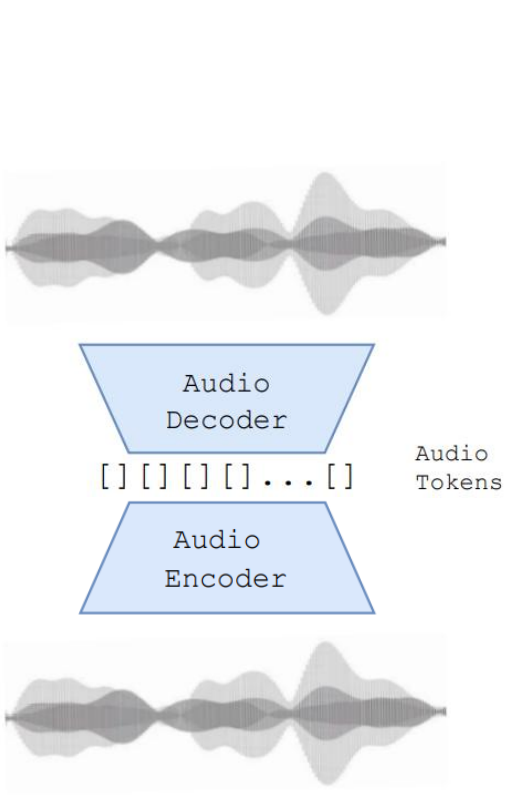
Audio Language Model



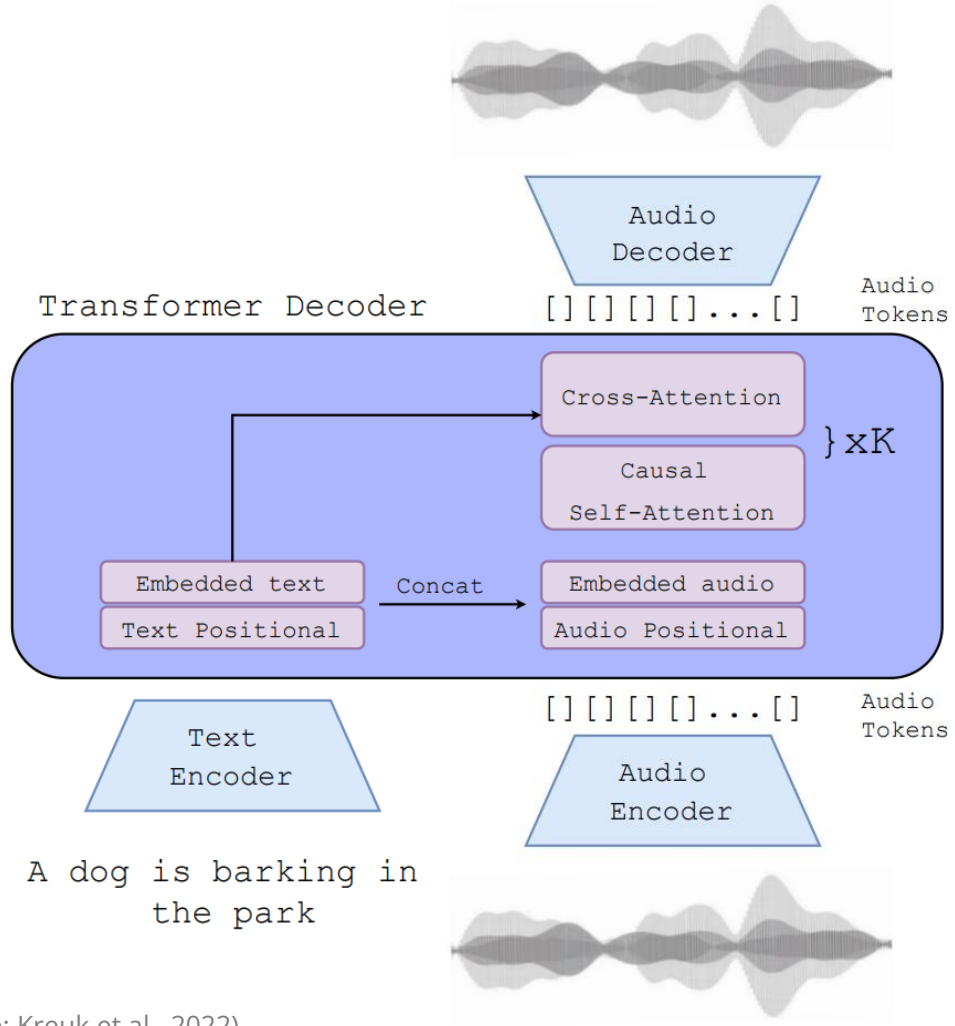
(Source: Kreuk et al., 2022)

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)

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)

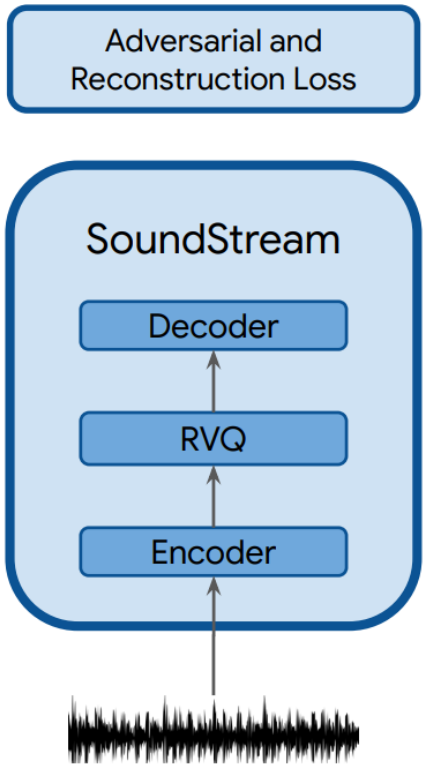
- AudioGen for Music
- Use EnCodec (Défossez et al., 2022) as the autoencoder
  - instead of SoundStream for AudioGen (Kreuk et al., 2023)
- **20k hours** of licensed music
  - Internal dataset 10k High-quality (private)
  - ShutterStock 25k Instrument-only
  - Pond5 365k Instrument-only

ai.honu.io/papers/musicgen/

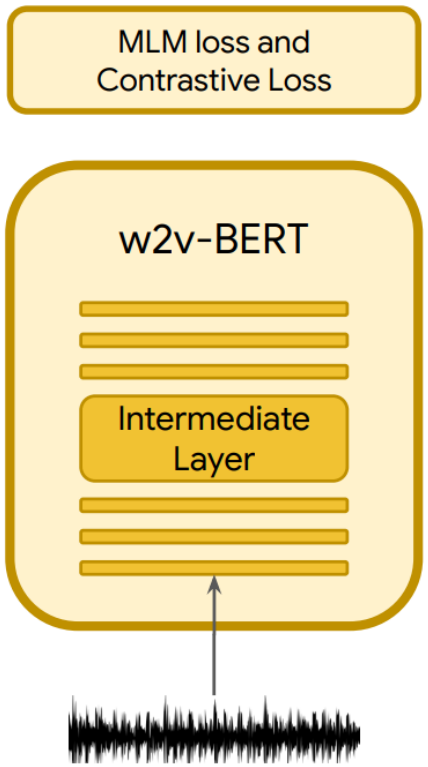
Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and Controllable Music Generation," *NeurIPS*, 2023.

MusicLM (Agostinelli et al., 2023)

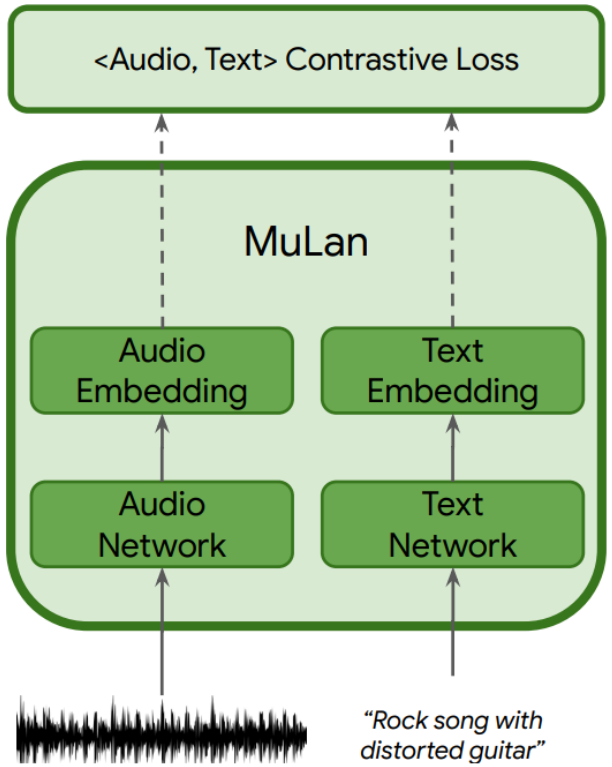
Audio autoencoder



Semantic representation



Text-music correspondence



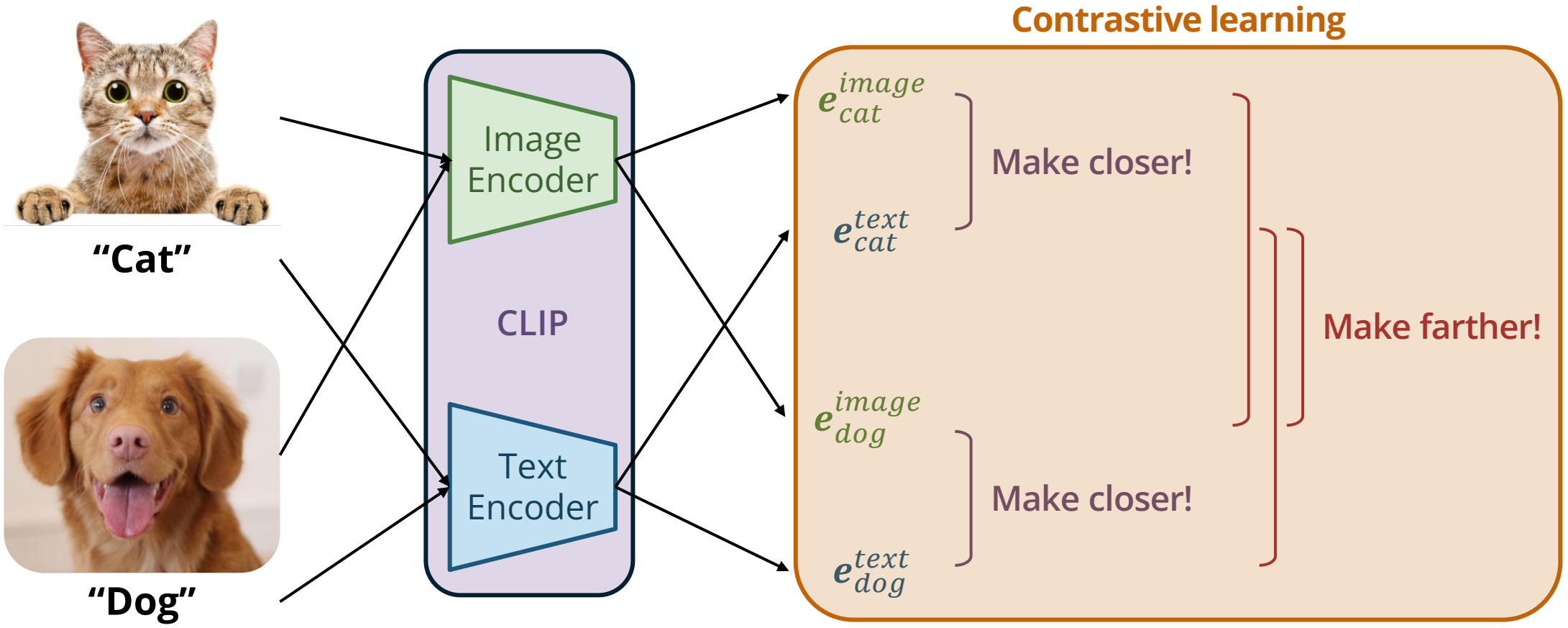
106k songs, 8.2k hours

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.

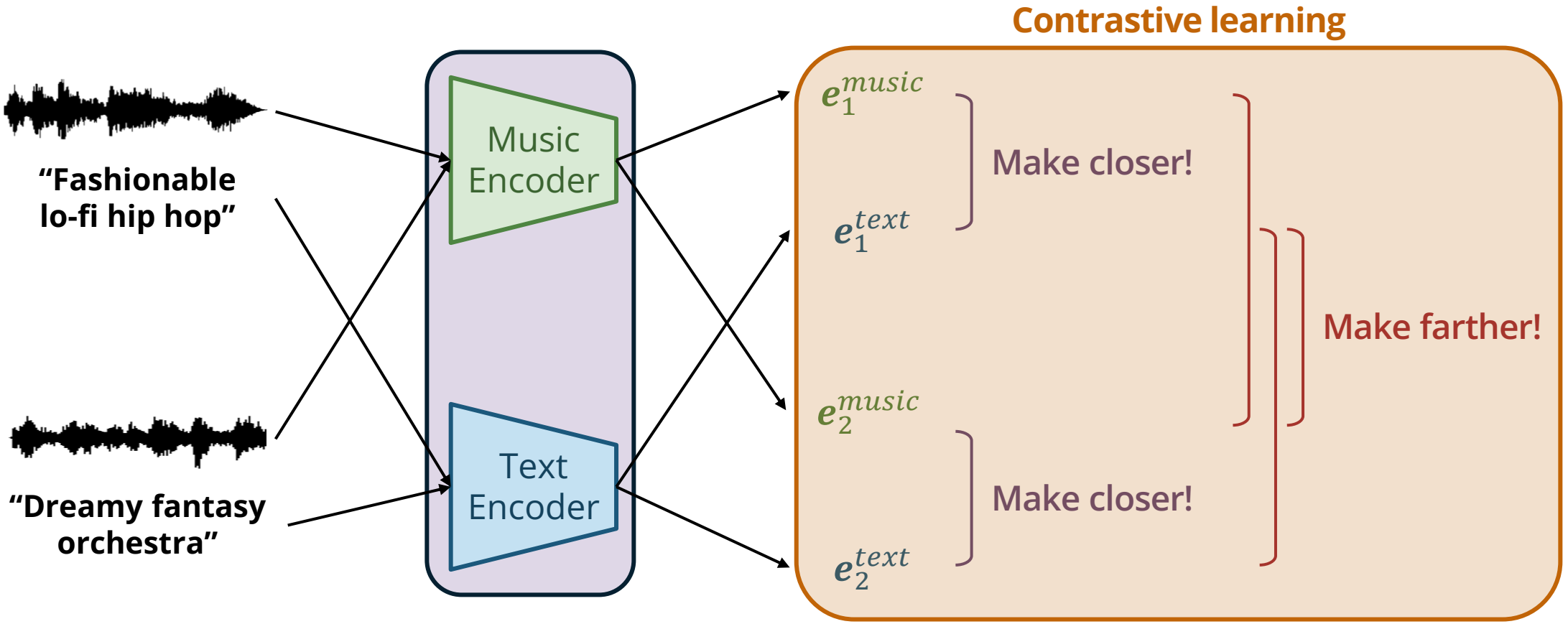
Contrastive Language-Image Pretraining (CLIP)



Learn a shared embedding space for images and language

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.

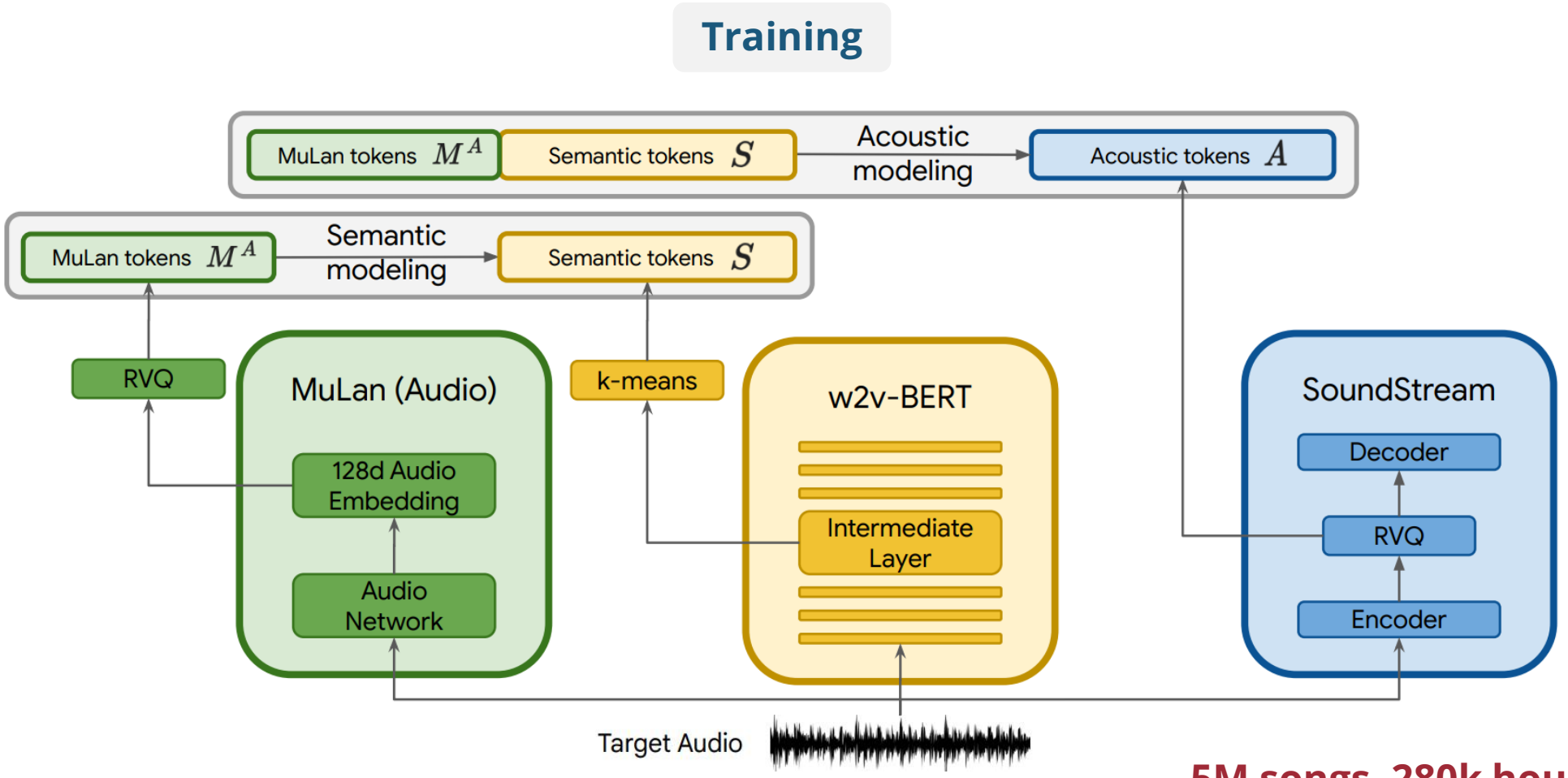
Contrastive Language-Music Pretraining



Learn a shared embedding space for music and language

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.

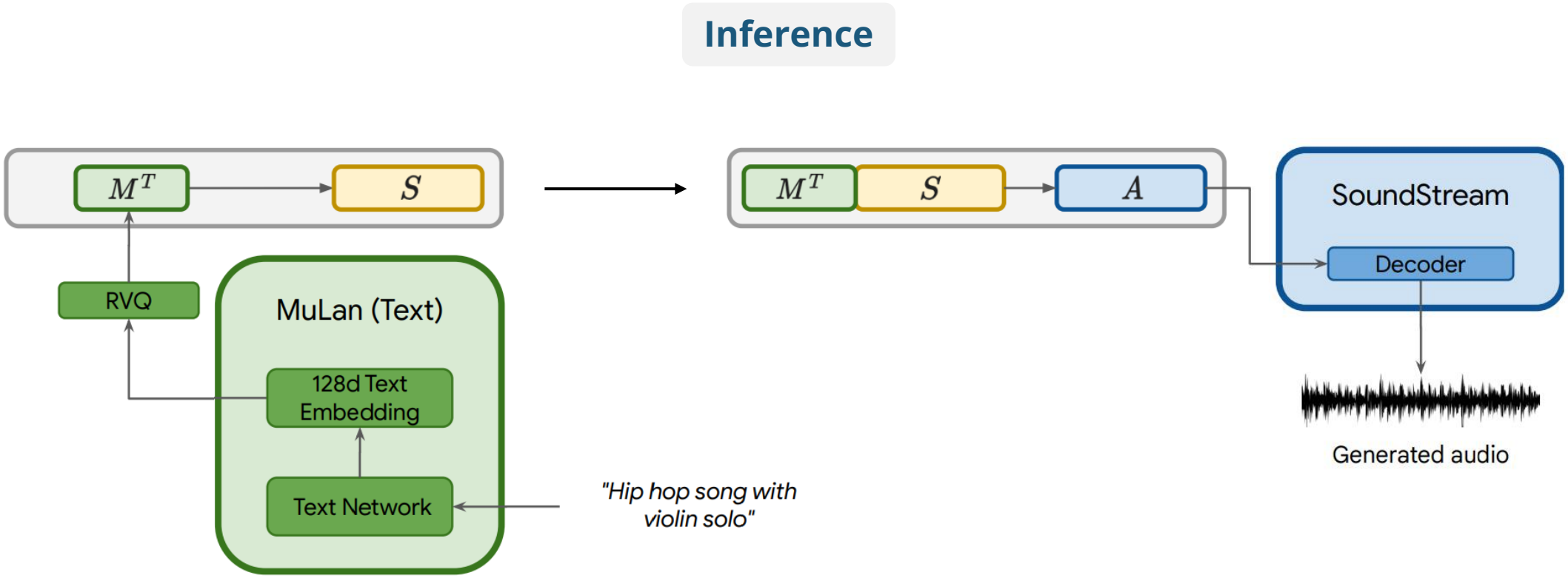
MusicLM (Agostinelli et al., 2023)



(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)



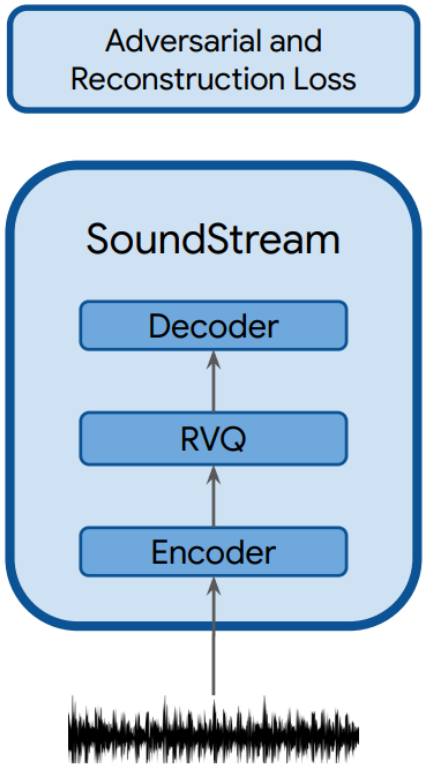
(Source: Agostinelli et al., 2022)

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

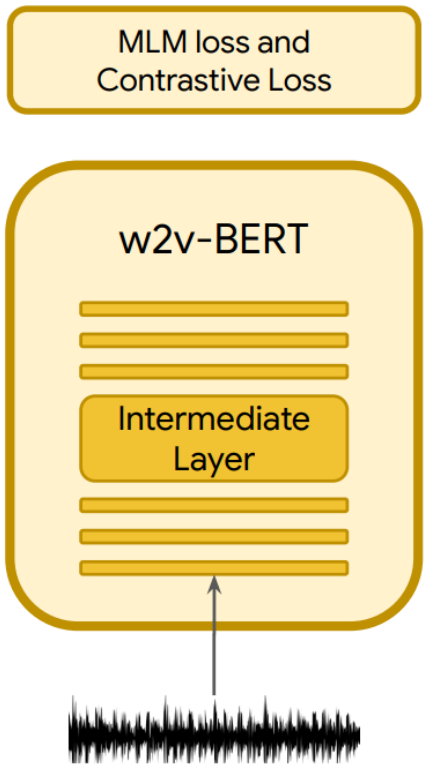
Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 2023.

MusicLM (Agostinelli et al., 2023)

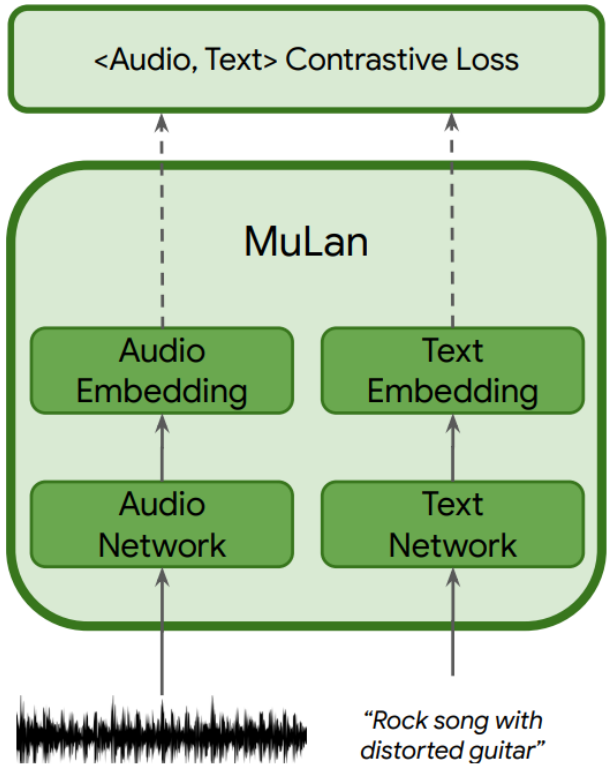
Audio autoencoder



Semantic representation



Text-music correspondence



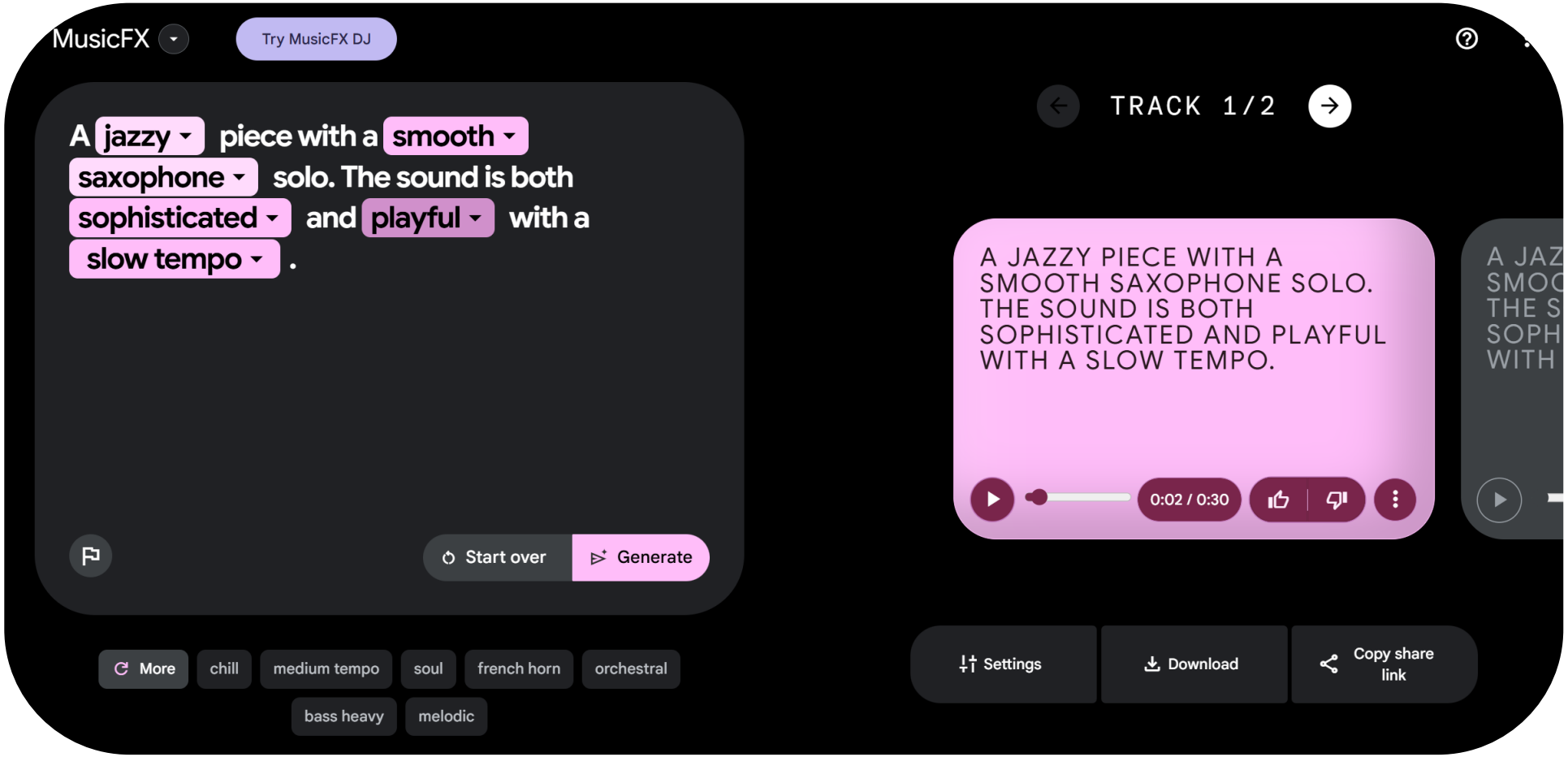
106k songs, 8.2k hours

44M 30-sec clips, 370k hours

(Source: Agostinelli et al., 2022)

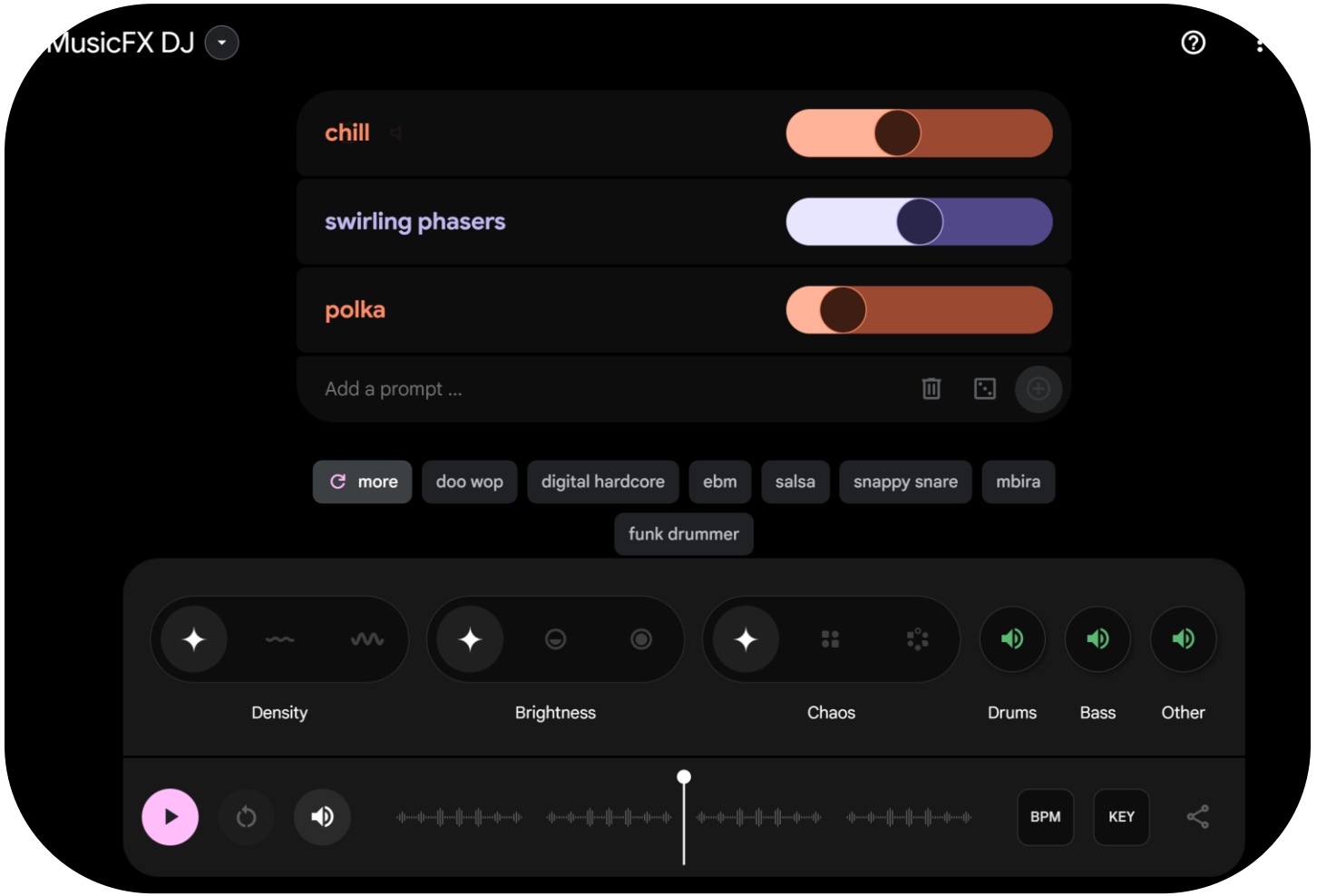
Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 2023.

Google's Music FX (2024)



labs.google/fx/tools/music-fx

Google's Music FX DJ (2024)

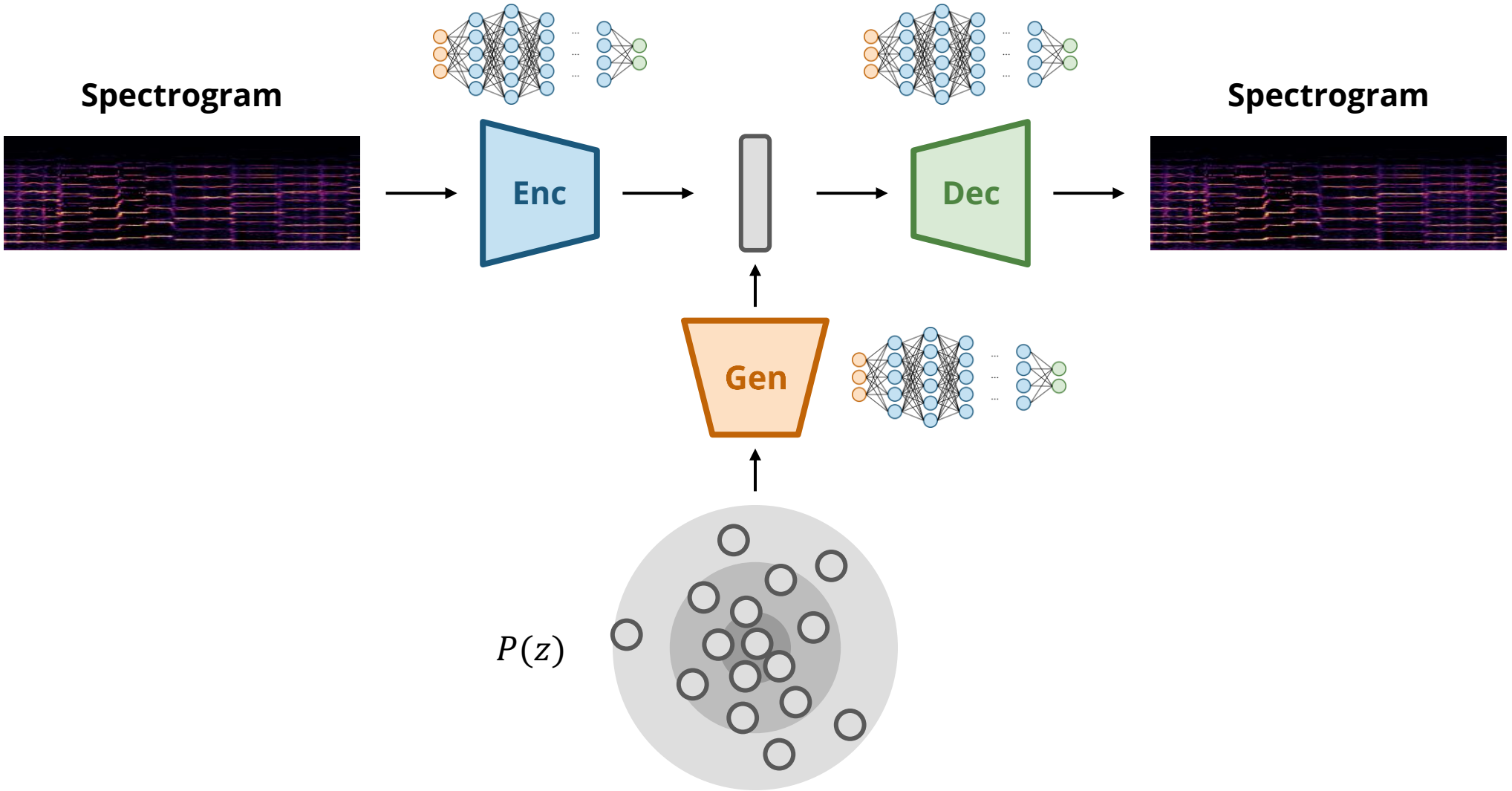


labs.google/fx/tools/music-fx-dj

Google's Music FX DJ (2024)

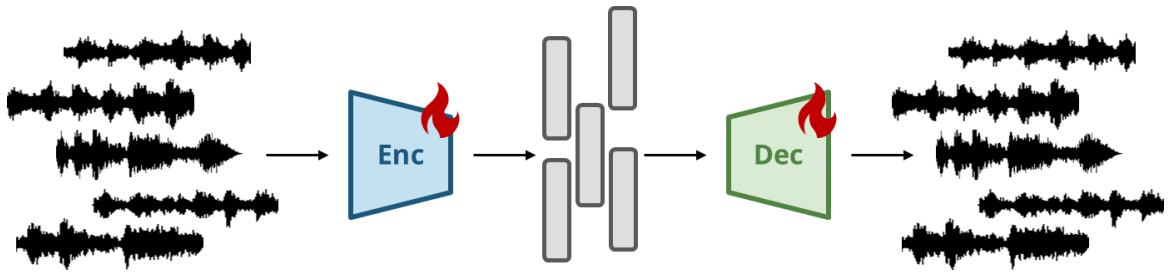
youtube.com/live/IUQW5LgBZvQ

Latent-based Audio Synthesis

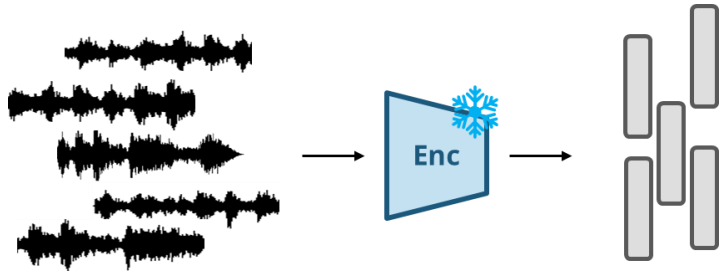


Latent-based Audio Synthesis: Pipeline

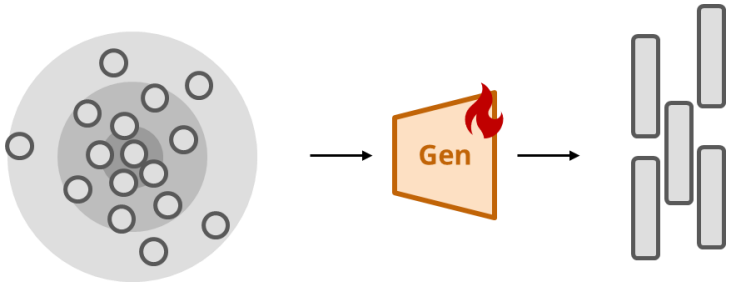
Step 1: Train an Autoencoder



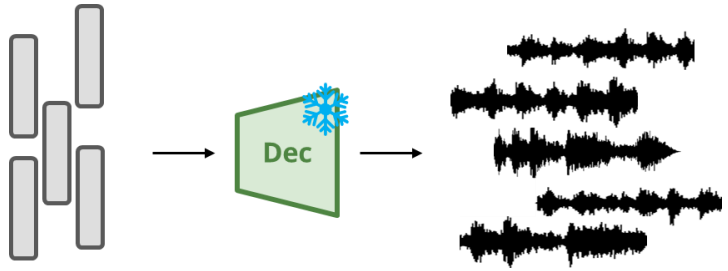
Step 2: Compute the Latent Vectors



Step 3: Train a Latent Generative Model

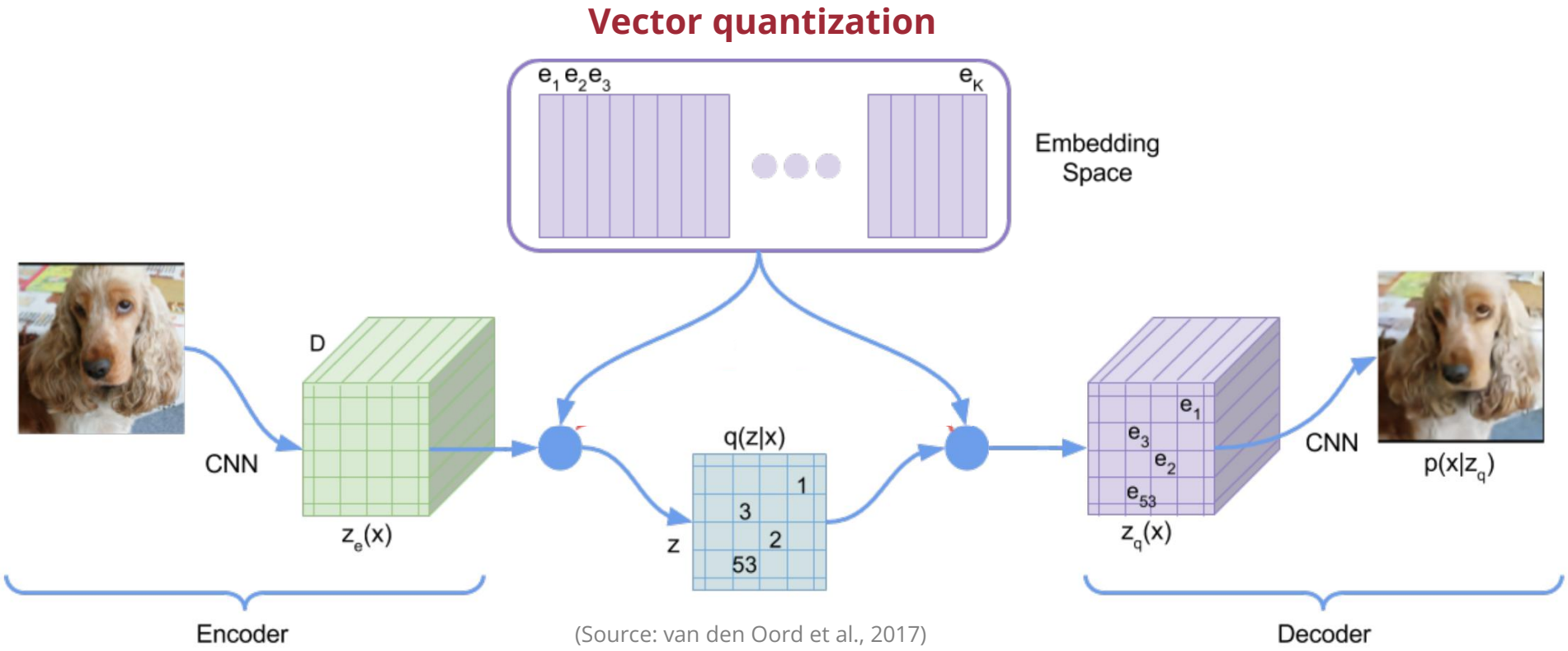


Step 4: Decode the Latent Vectors



Latent Diffusion Models

Vector-Quantized VAE (VQVAE)

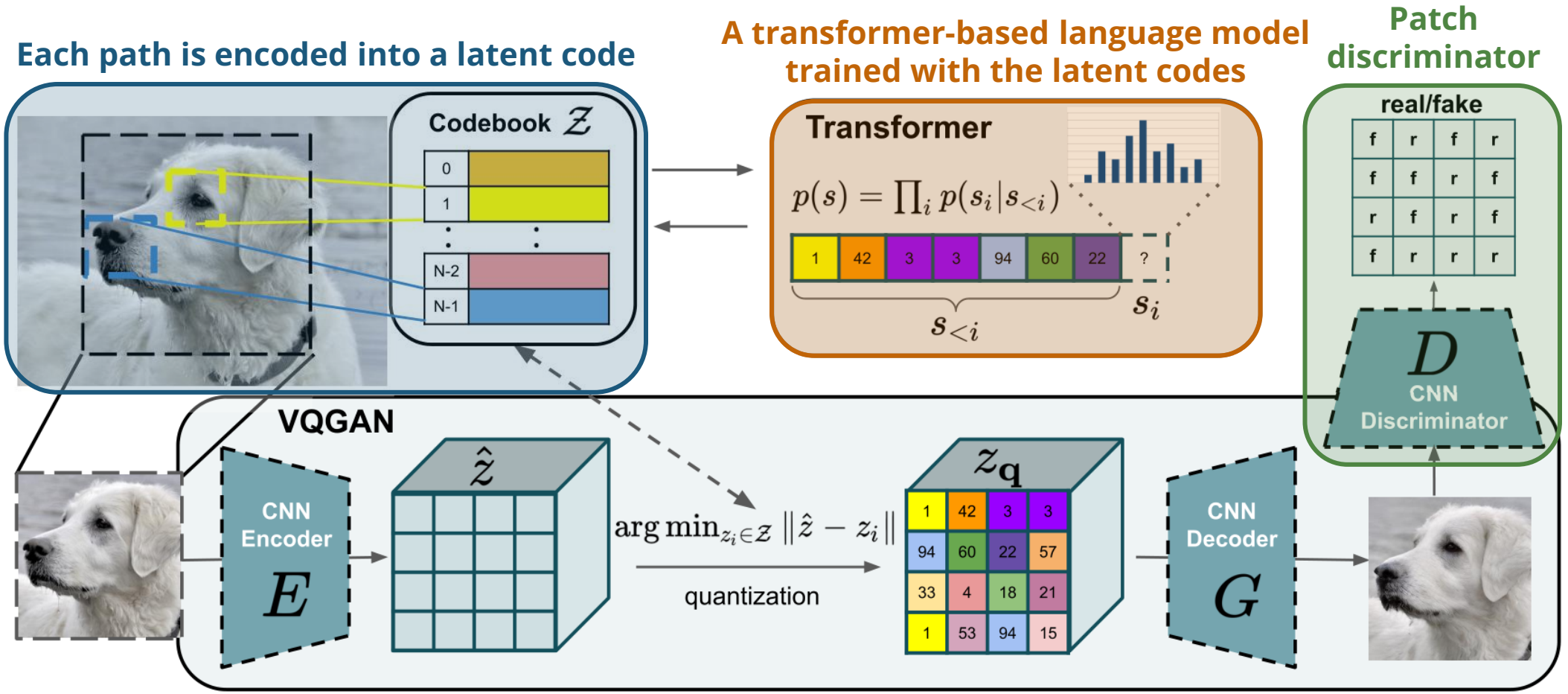


(Source: van den Oord et al., 2017)

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.

Vector-Quantized GAN (VQGAN)

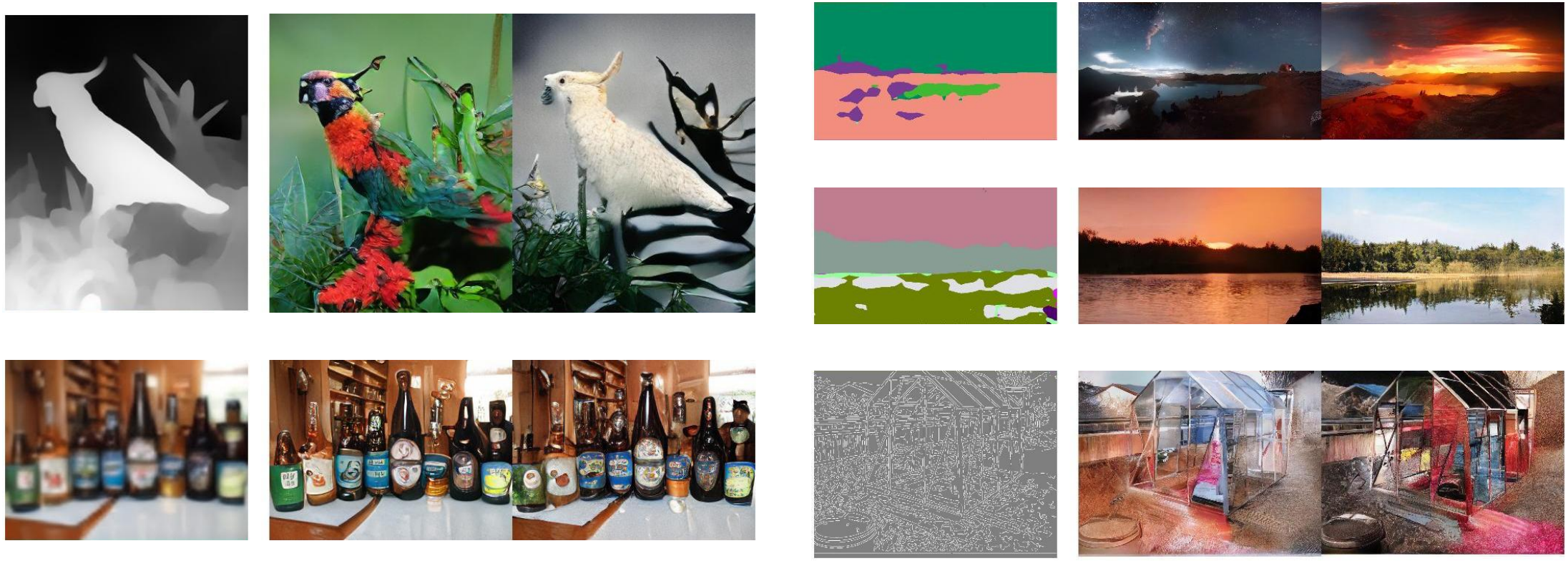


(Source: Esser et al., 2021)

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.

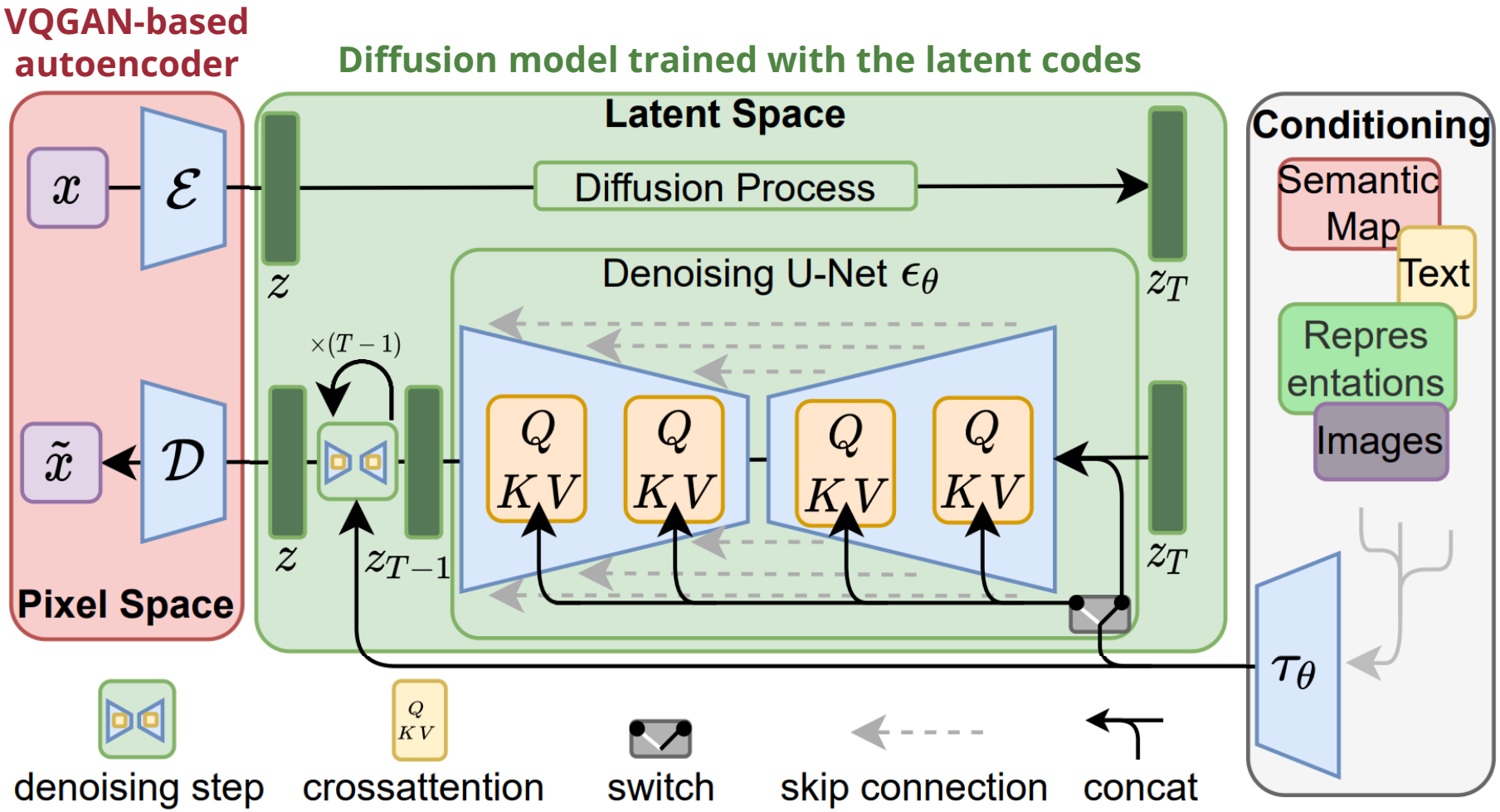
VQGAN: Conditional Generation



(Source: Esser et al., 2021)

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

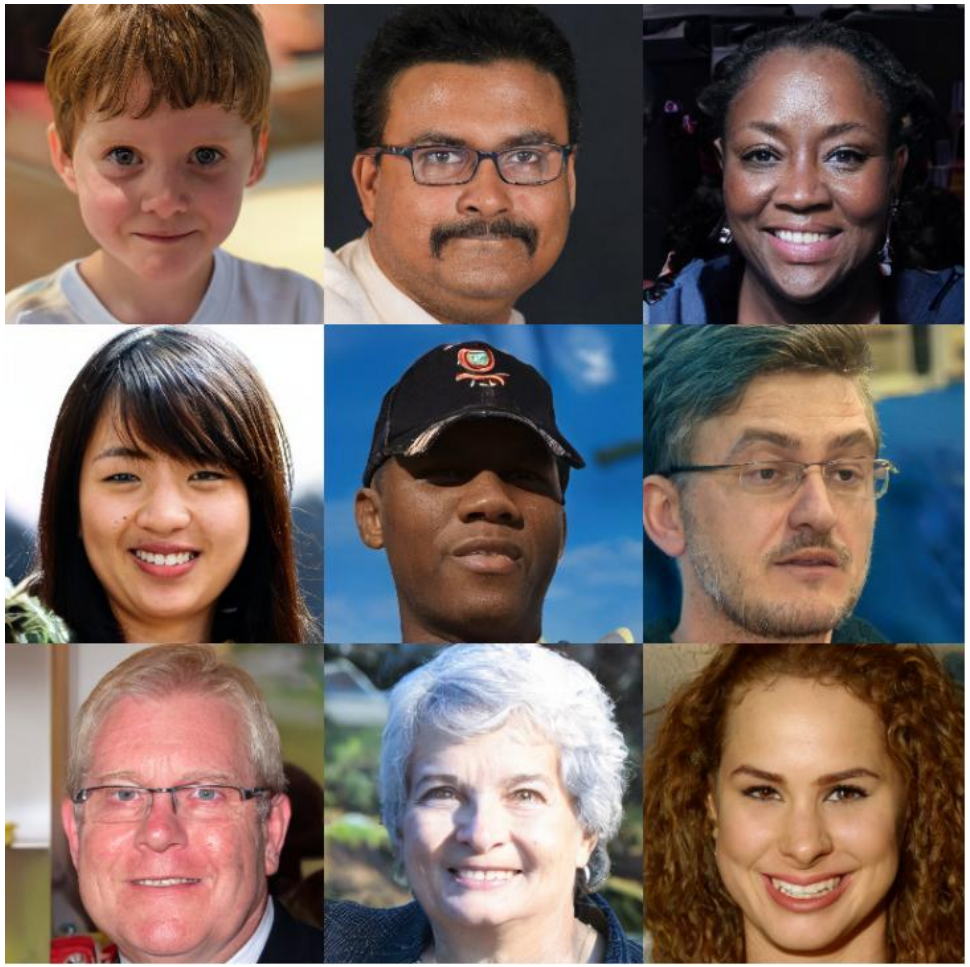
Latent Diffusion Models (LDMs)



(Source: Rombach et al., 2022)

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

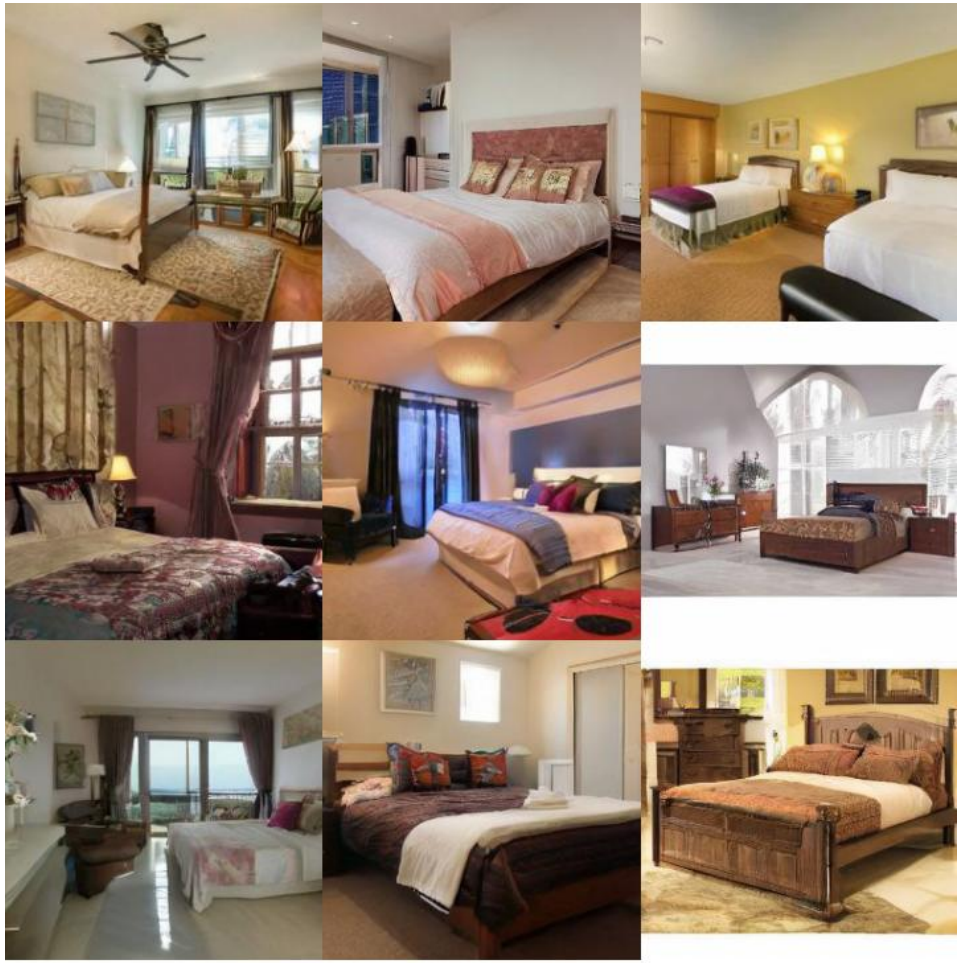
Latent Diffusion Models: Unconditional Generation



(Source: Rombach et al., 2022)

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

Latent Diffusion Models: Unconditional Generation

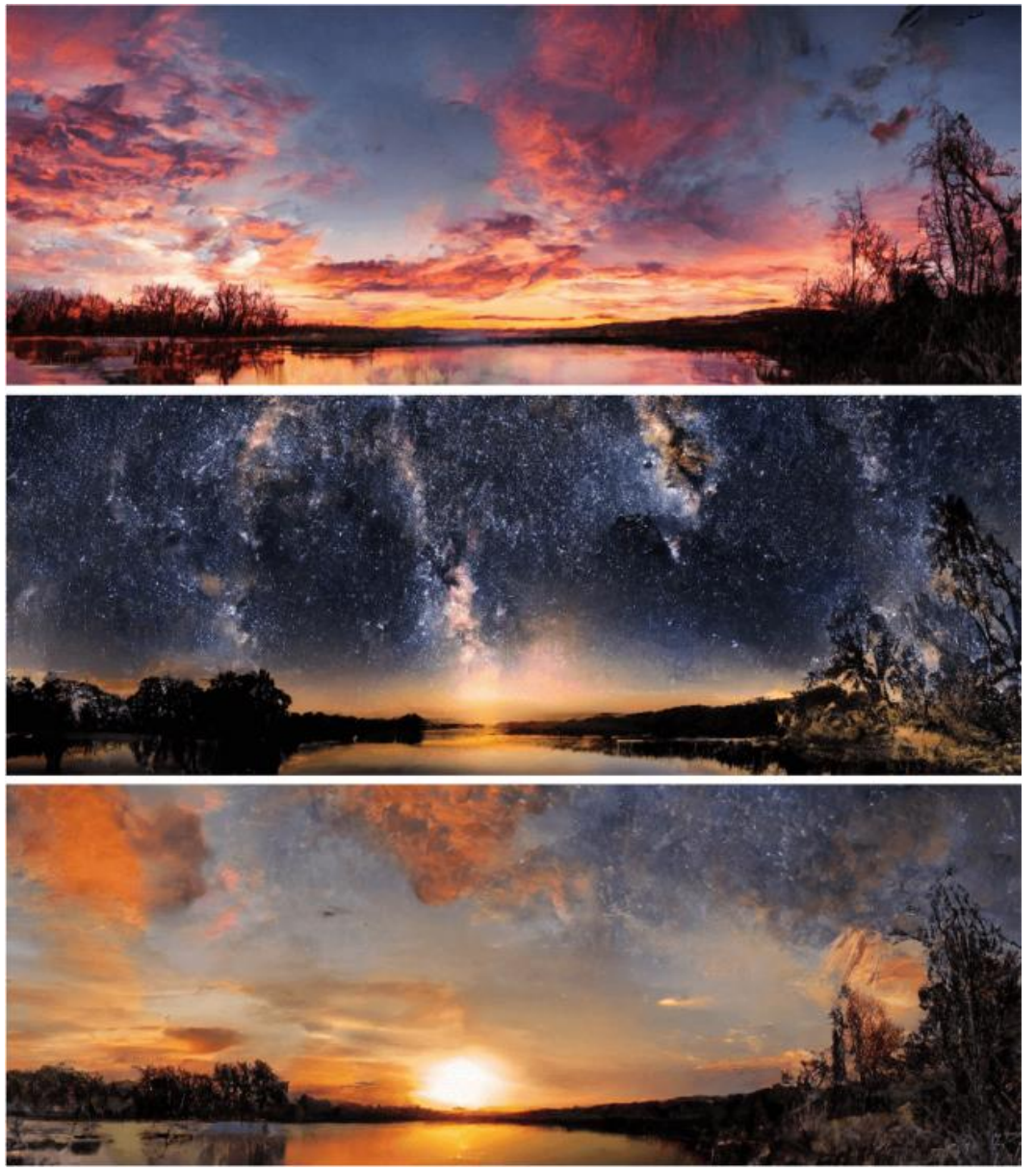


(Source: Rombach et al., 2022)

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

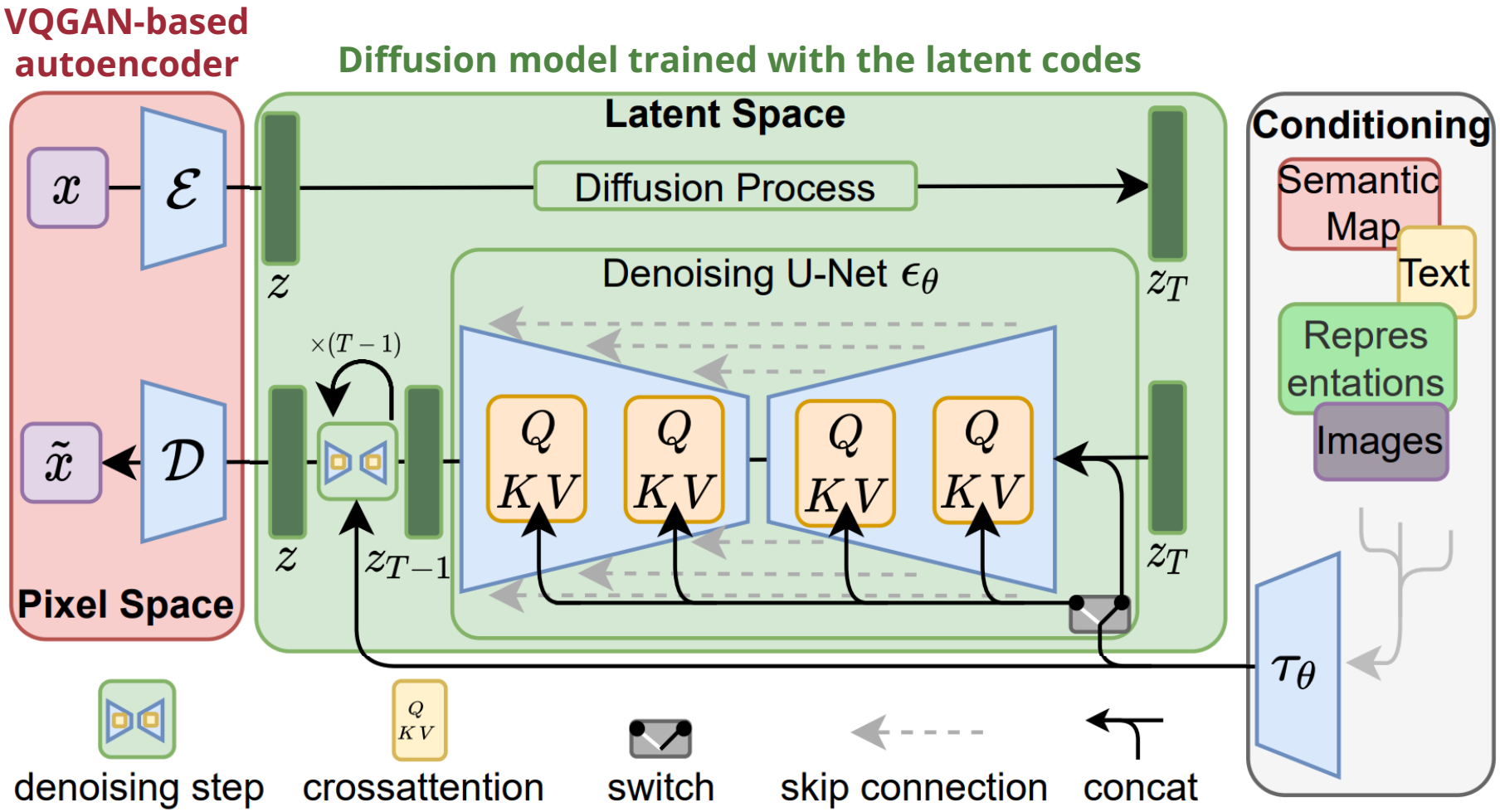
LDMs: Semantic Synthesis

(Source: Rombach et al., 2022)



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

Latent Diffusion Models (LDMs)

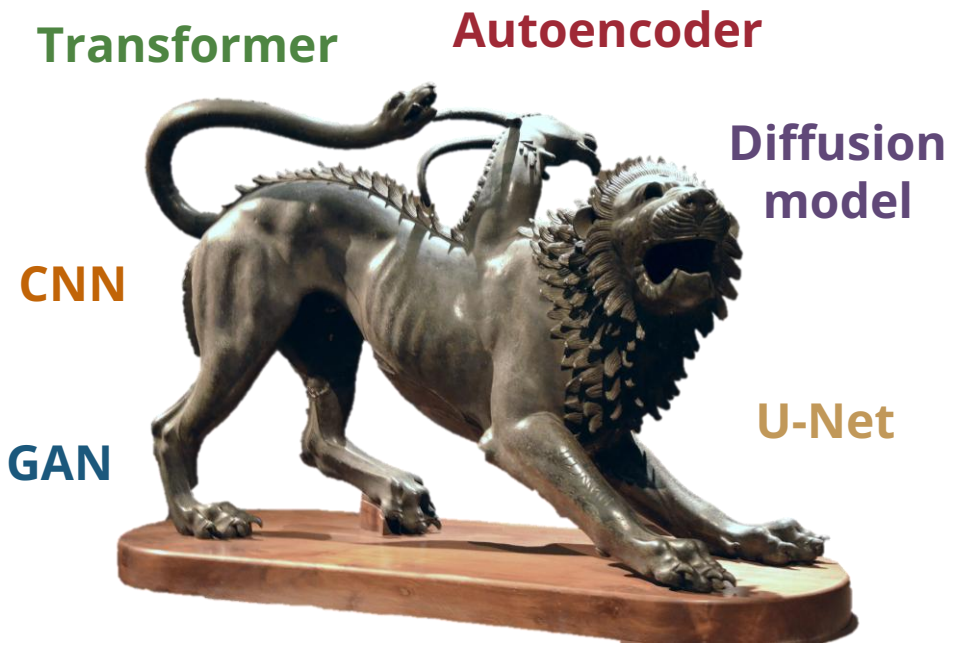


(Source: Rombach et al., 2022)

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

Latent Diffusion Model is a Chimera

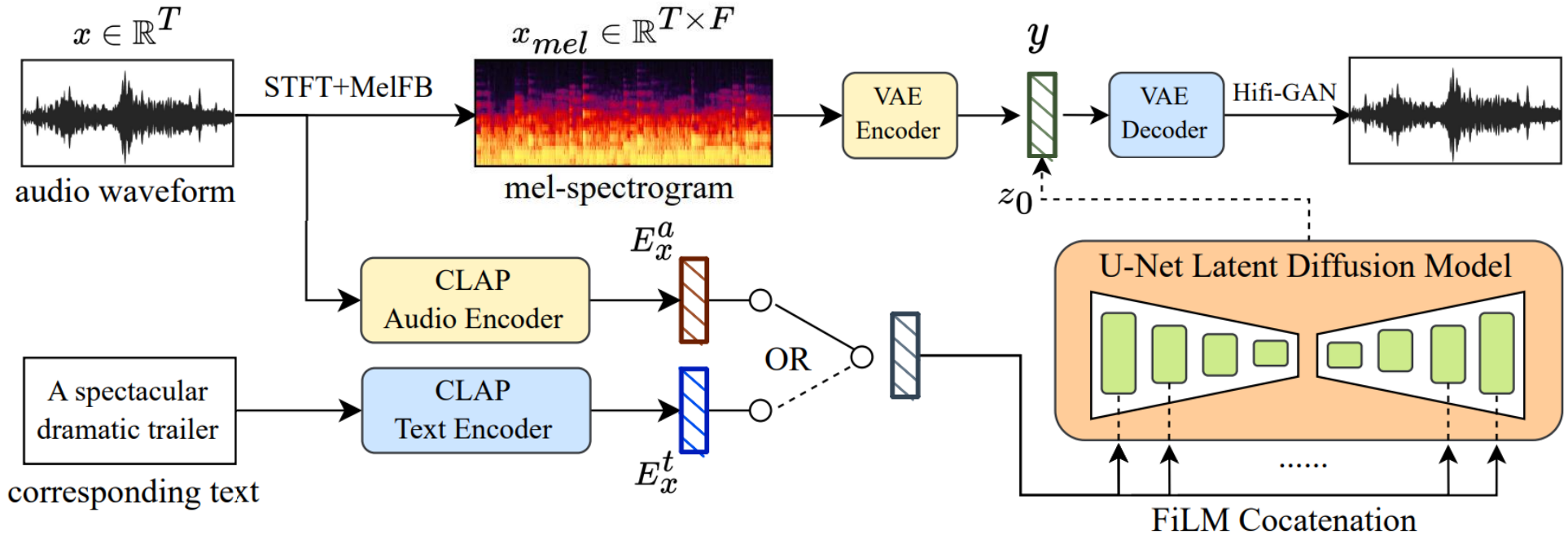
- **A neural codec**
  - An CNN-based autoencoder
  - Trained with a GAN-like adversarial loss
- **Diffusion model in the latent space**
  - A denoising U-Net
- **A conditioning module**
  - Transformer-like cross-attention mechanism



(Source: Raddato via worldhistory.org)

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

Latent Diffusion for Music: MusicLDM (Chen et al., 2023)

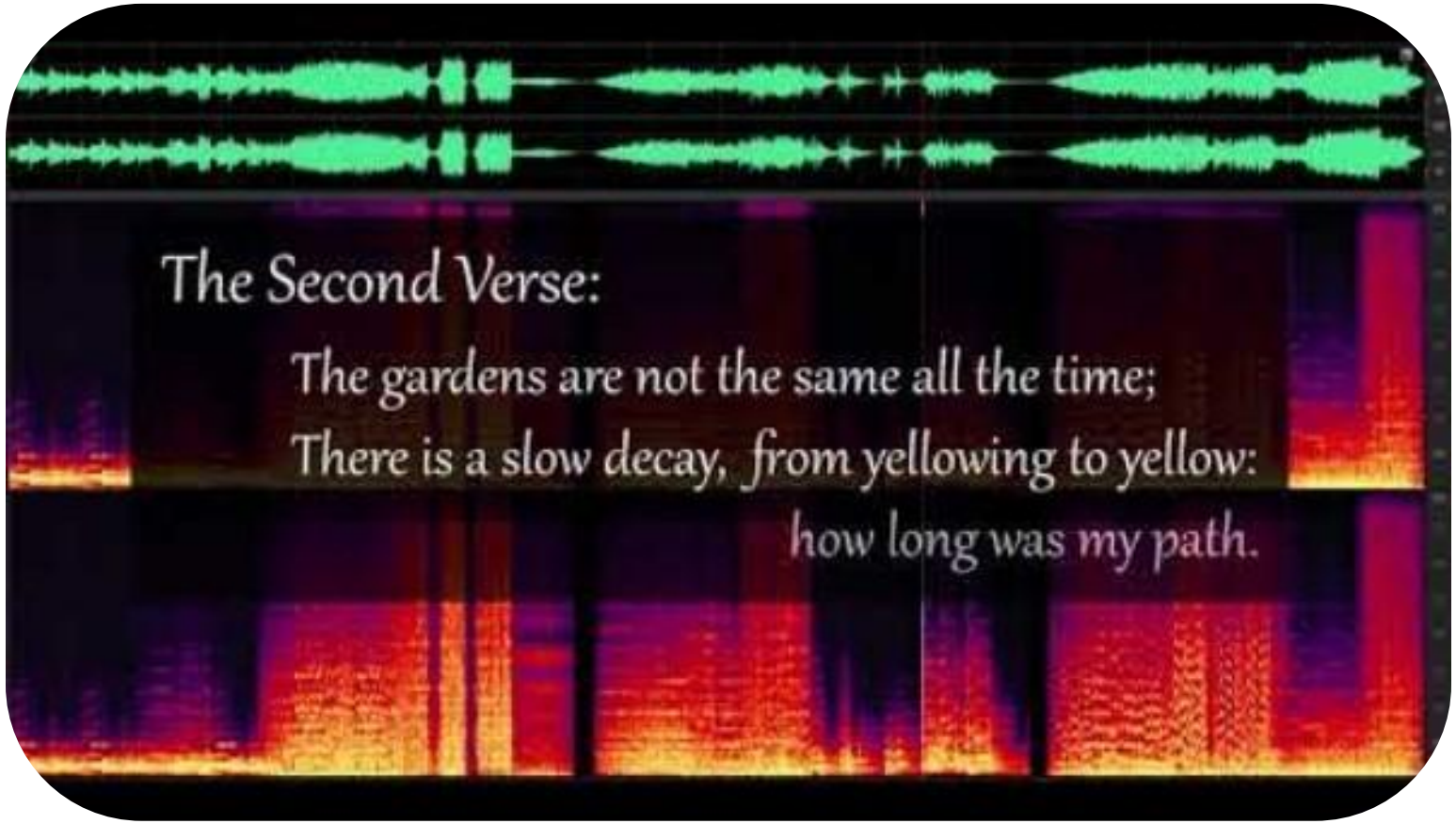


(Source: Ke et al., 2023)

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.

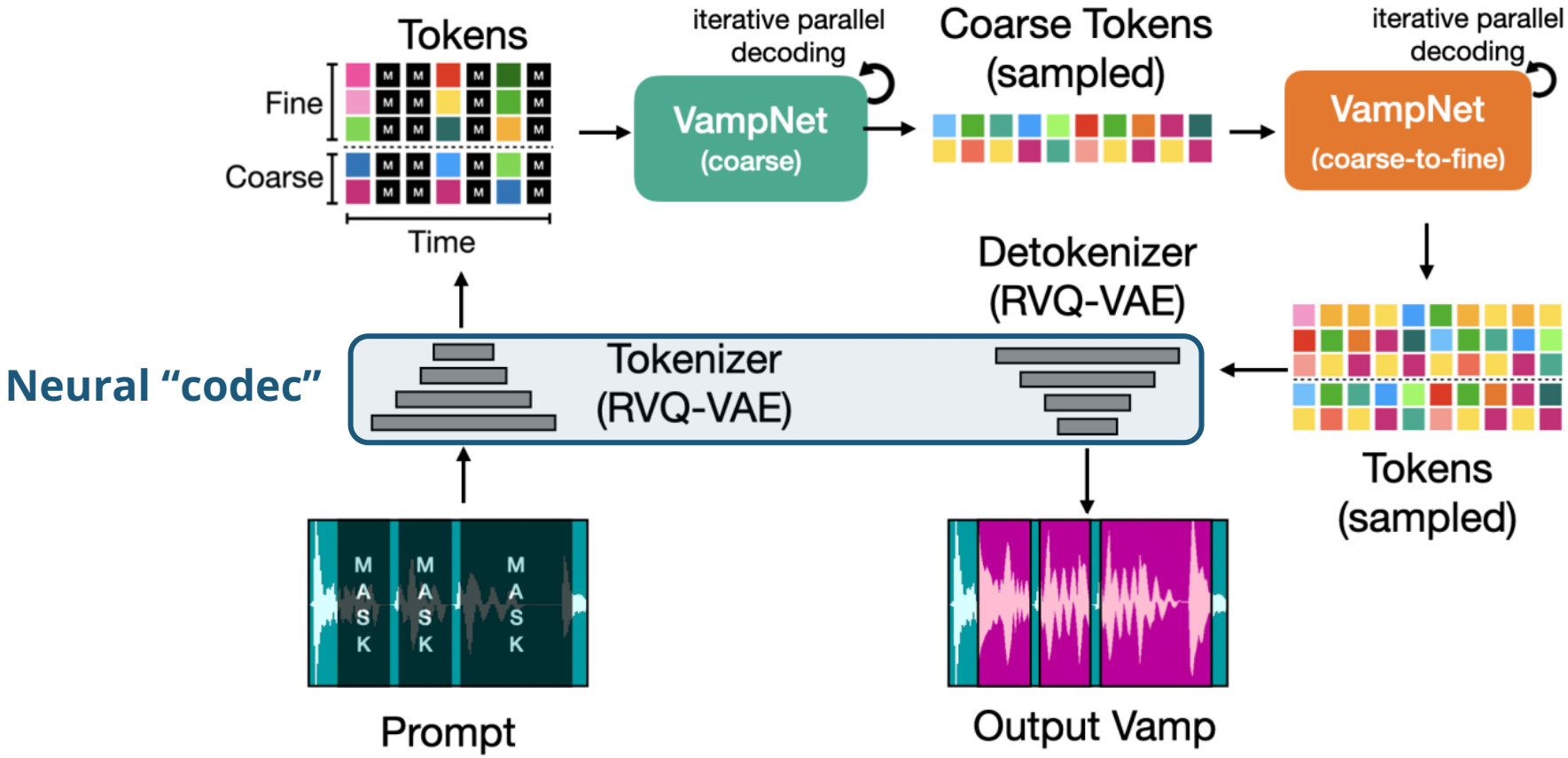
MusicLDM: Demo (Chen et al., 2023)



youtu.be/DALv7ea6cv0

Masked Acoustic Token Modeling

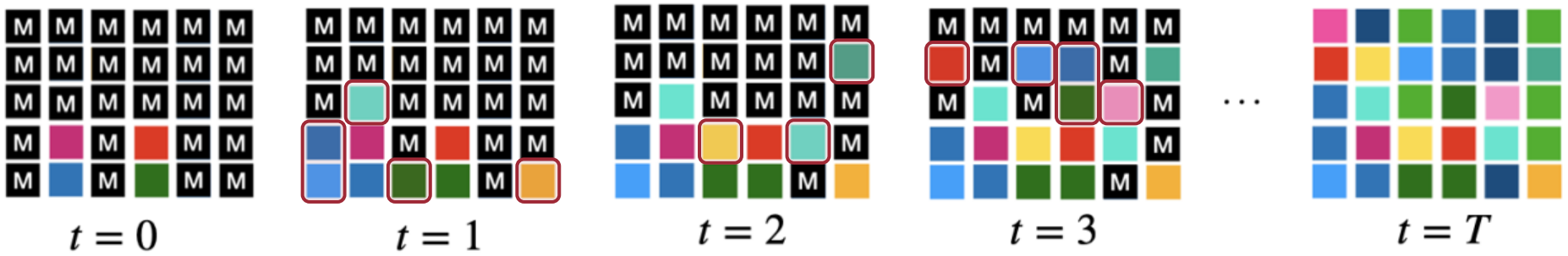
VampNet (Garcia et al., 2023)



(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)

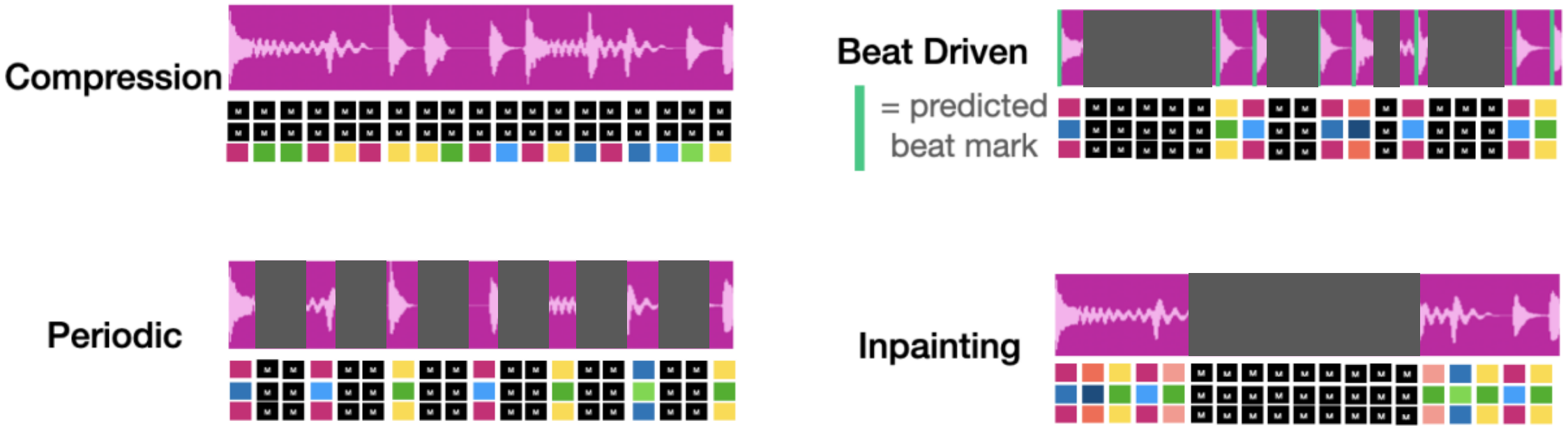


(Source: Garcia et al., 2023)

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.

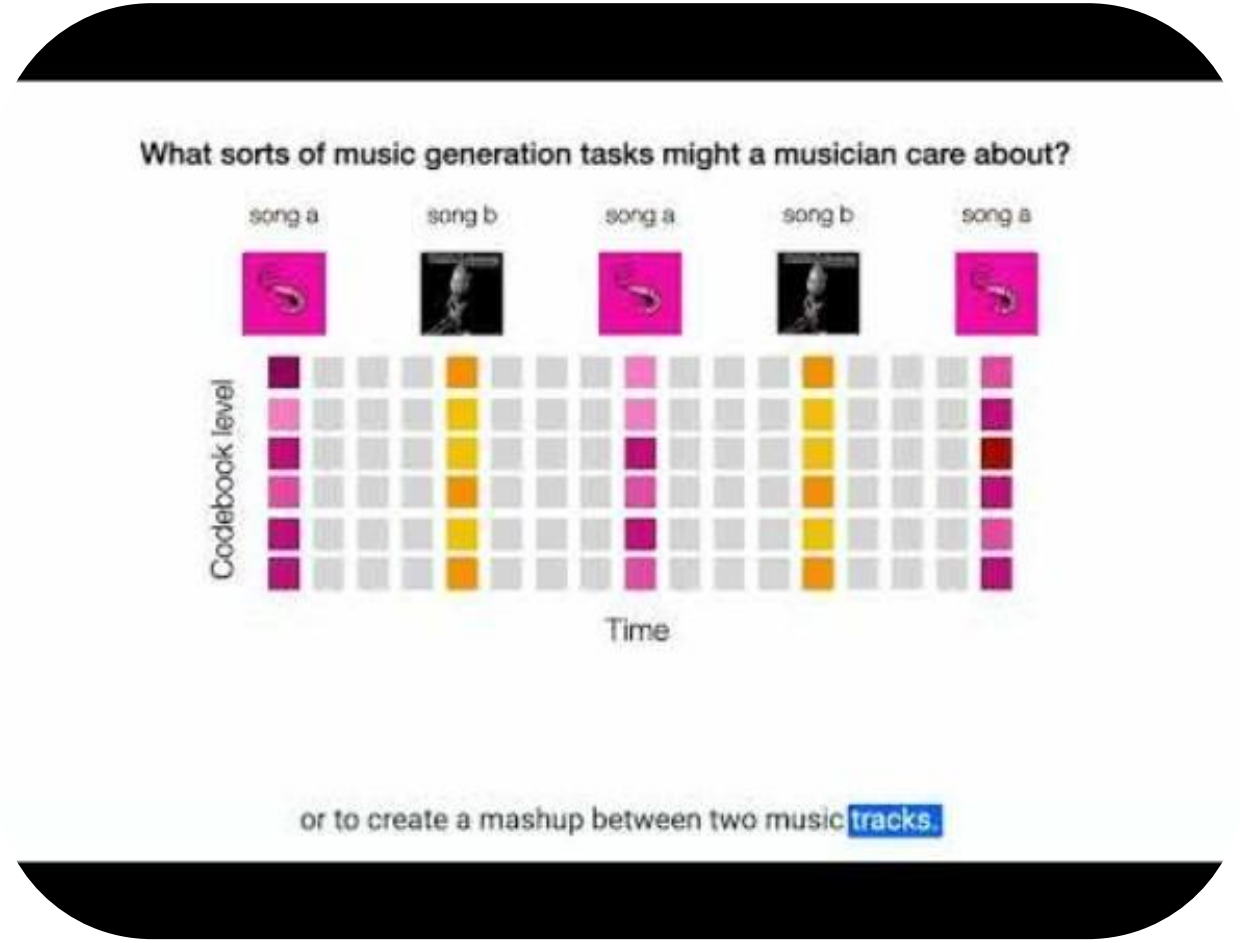
VampNet (Garcia et al., 2023)



(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

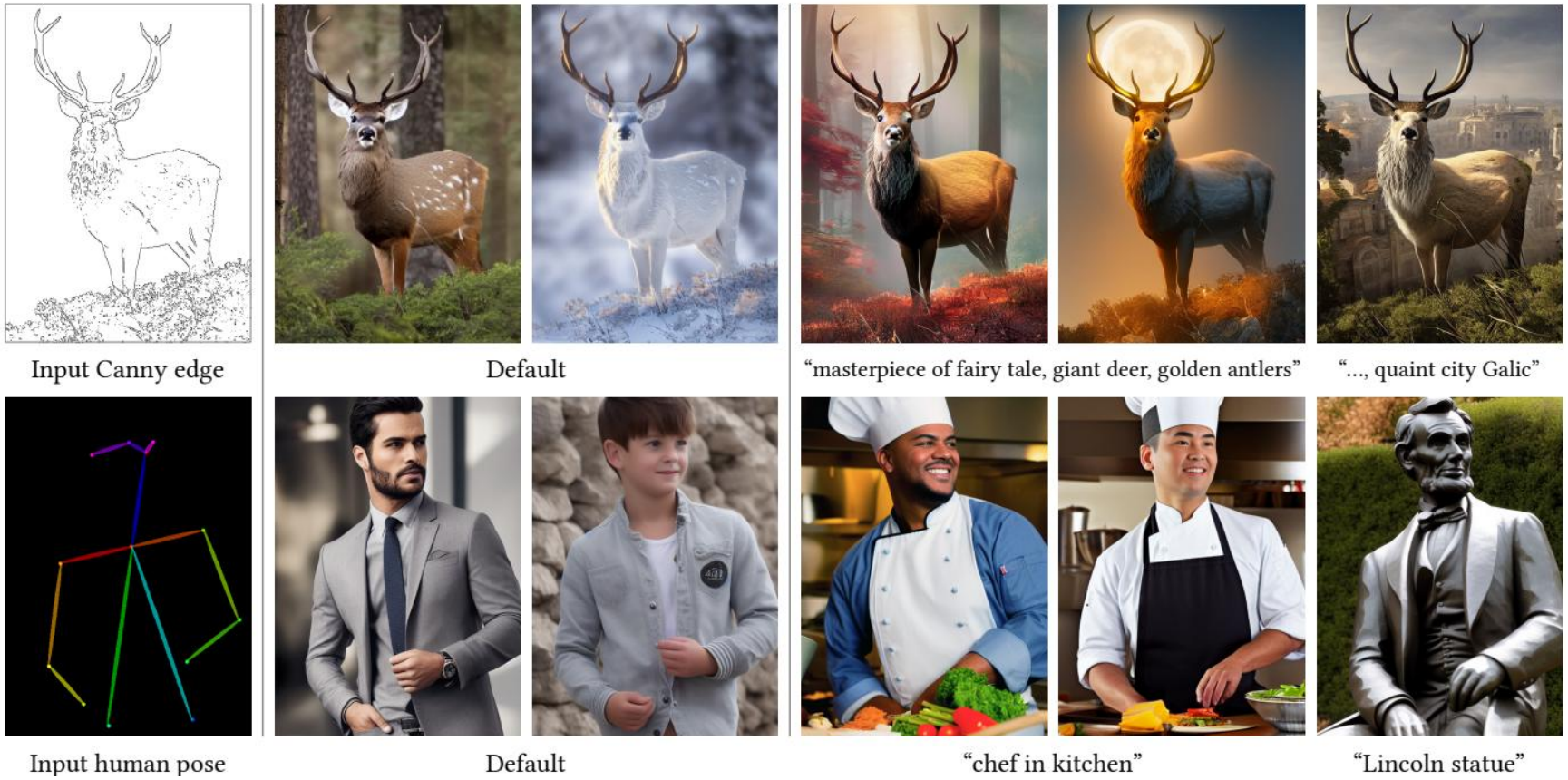
Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo, "VampNet: Music Generation via Masked Acoustic Token Modeling," *ISMIR*, 2023.

`unloop` (Garcia et al., 2023)

youtu.be/yzBI8Vcjd2s & github.com/hugofloresgarcia/unloop

Controlling Music Generation Systems

ControlNet (Zhang et al., 2023)

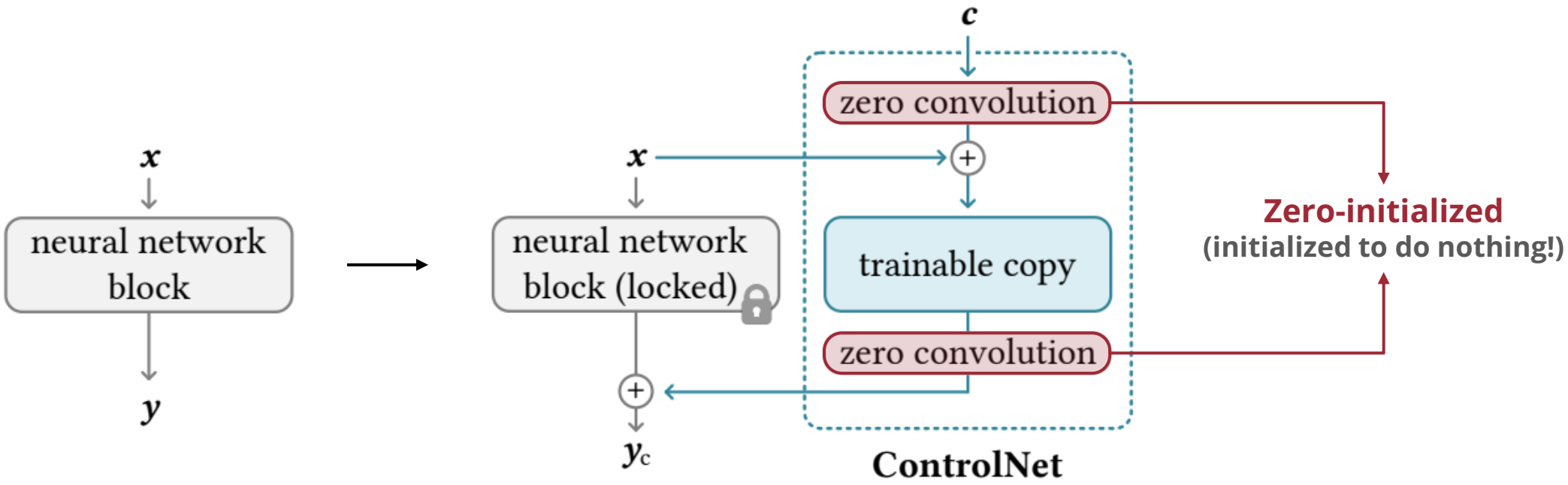


(Source: Zhang et al., 2023)

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.

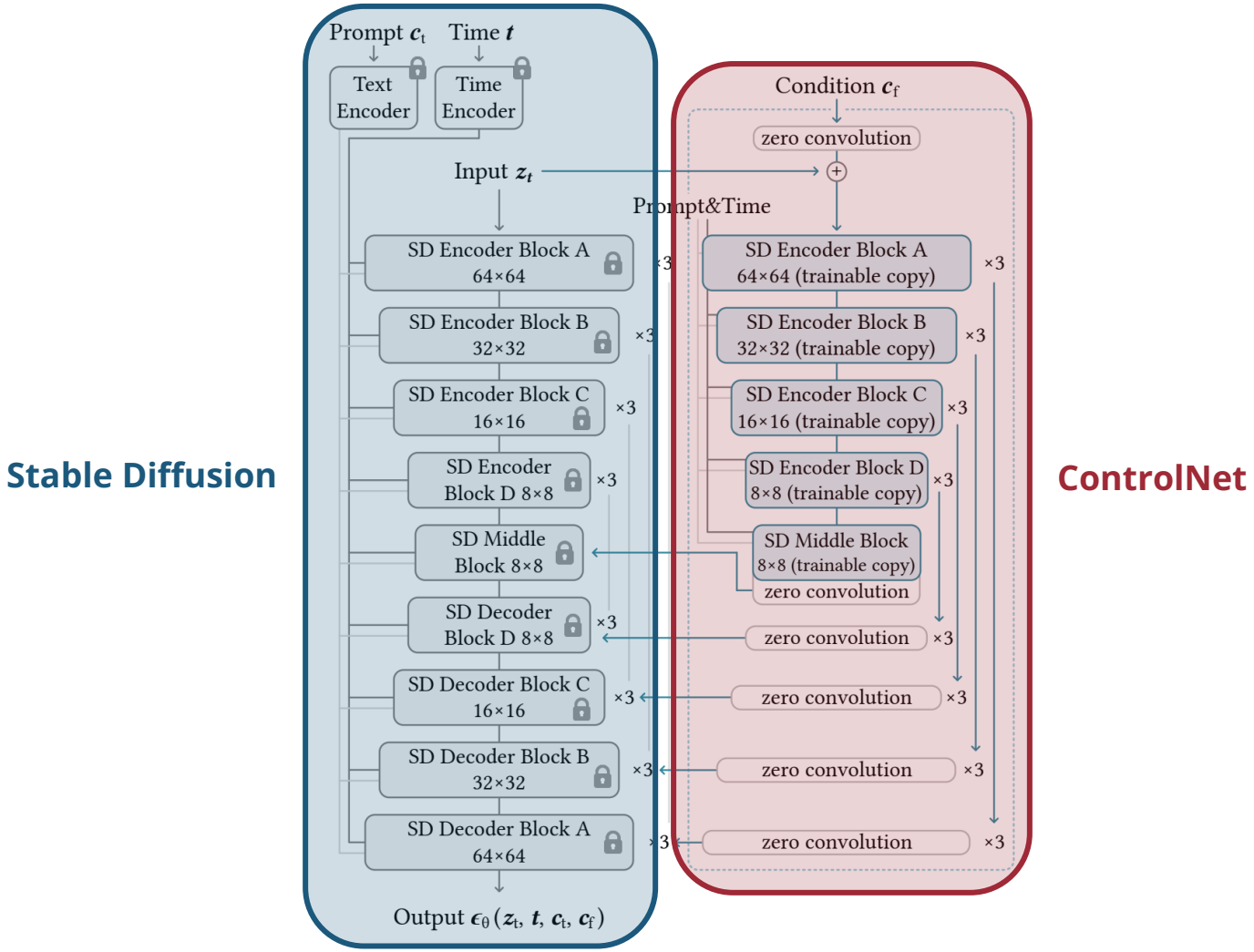
ControlNet (Zhang et al., 2023)



(Source: Zhang et al., 2023)

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

ControlNet (Zhang et al., 2023)



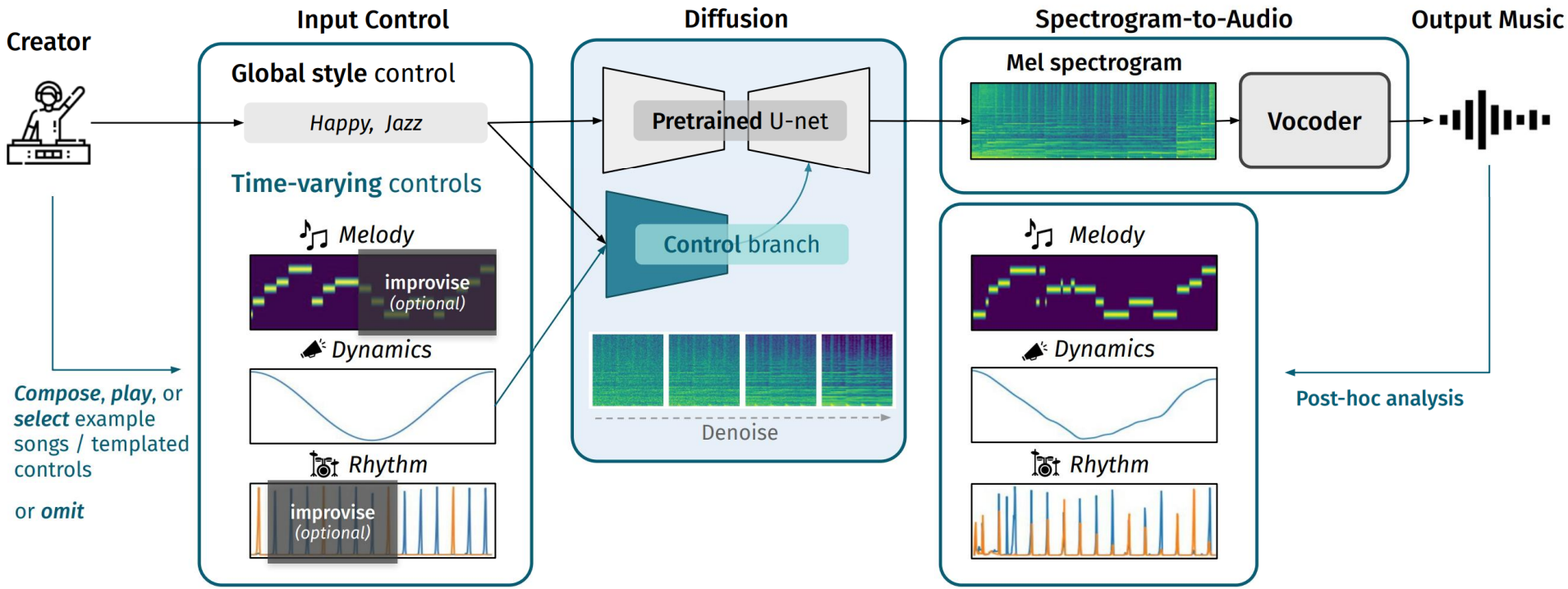
(Source: Zhang et al., 2023)

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

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

youtu.be/O4cJ3acEGDw

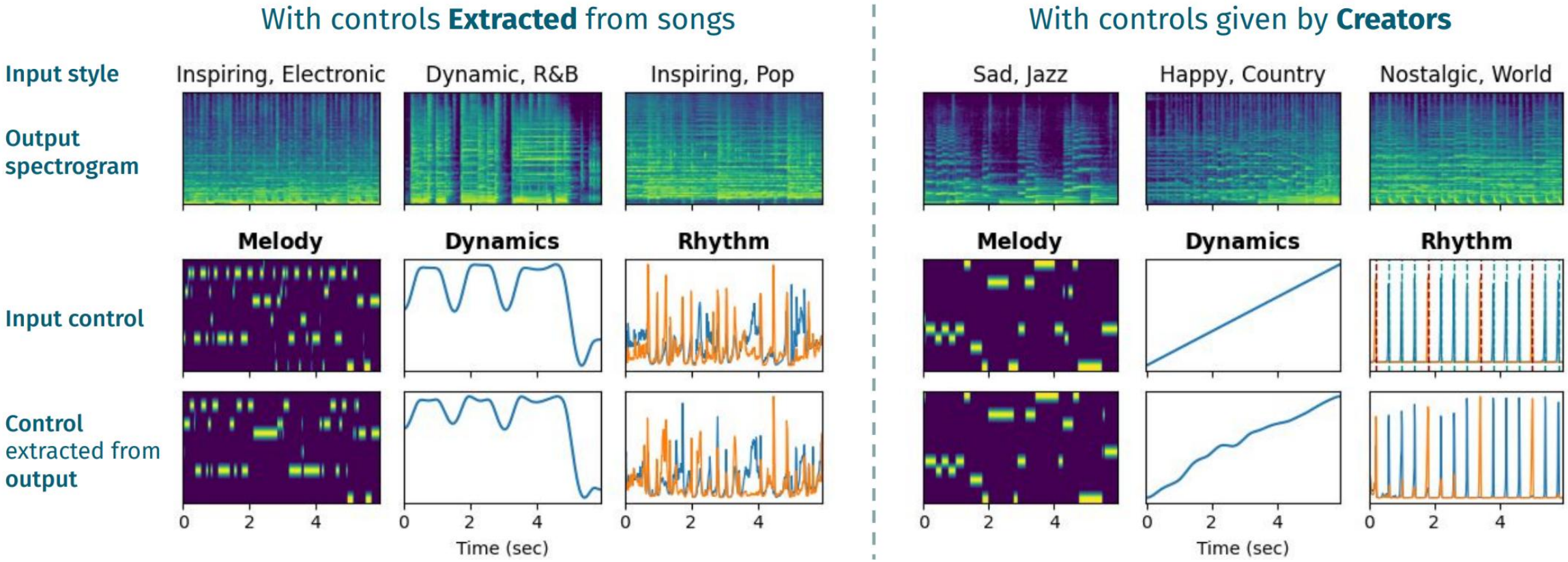
Music ControlNet (Wu et al., 2024)



(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)

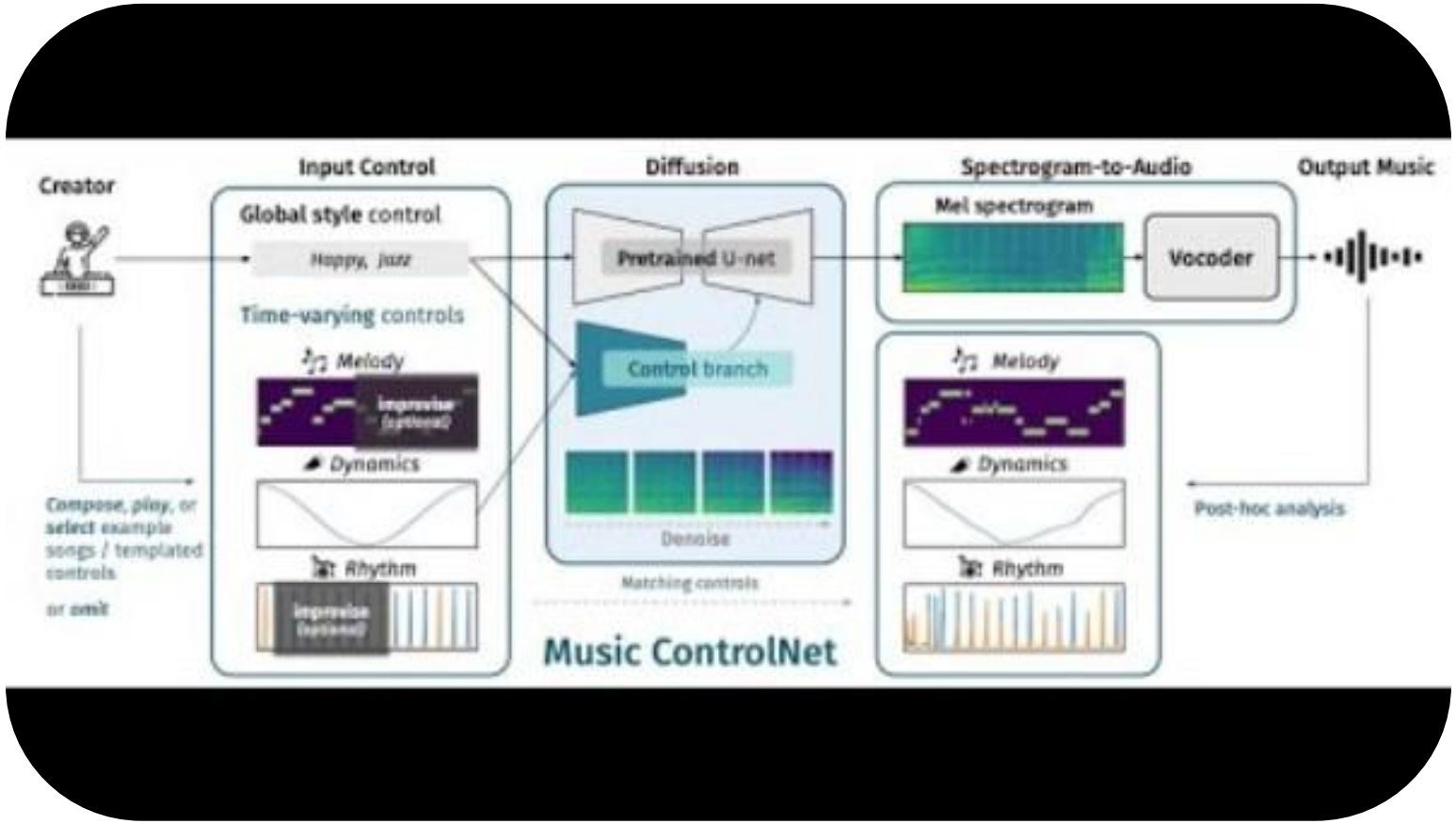


(Source: Wu et al., 2024)

musiccontrolnet.github.io/web

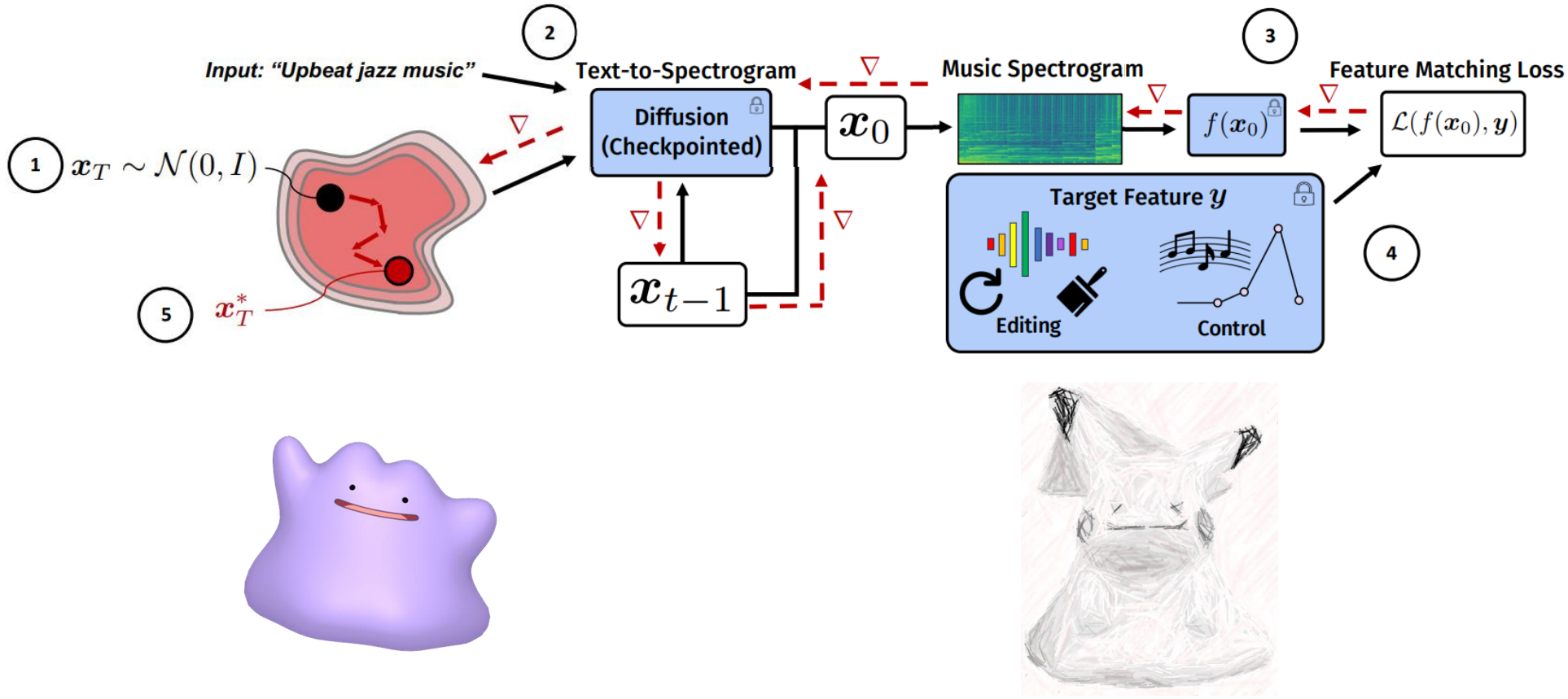
Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan, "Music ControlNet: Multiple Time-varying Controls for Music Generation," *TASLP*, 2024.

Music ControlNet (Wu et al., 2024)



youtu.be/QVr-S-DyccU

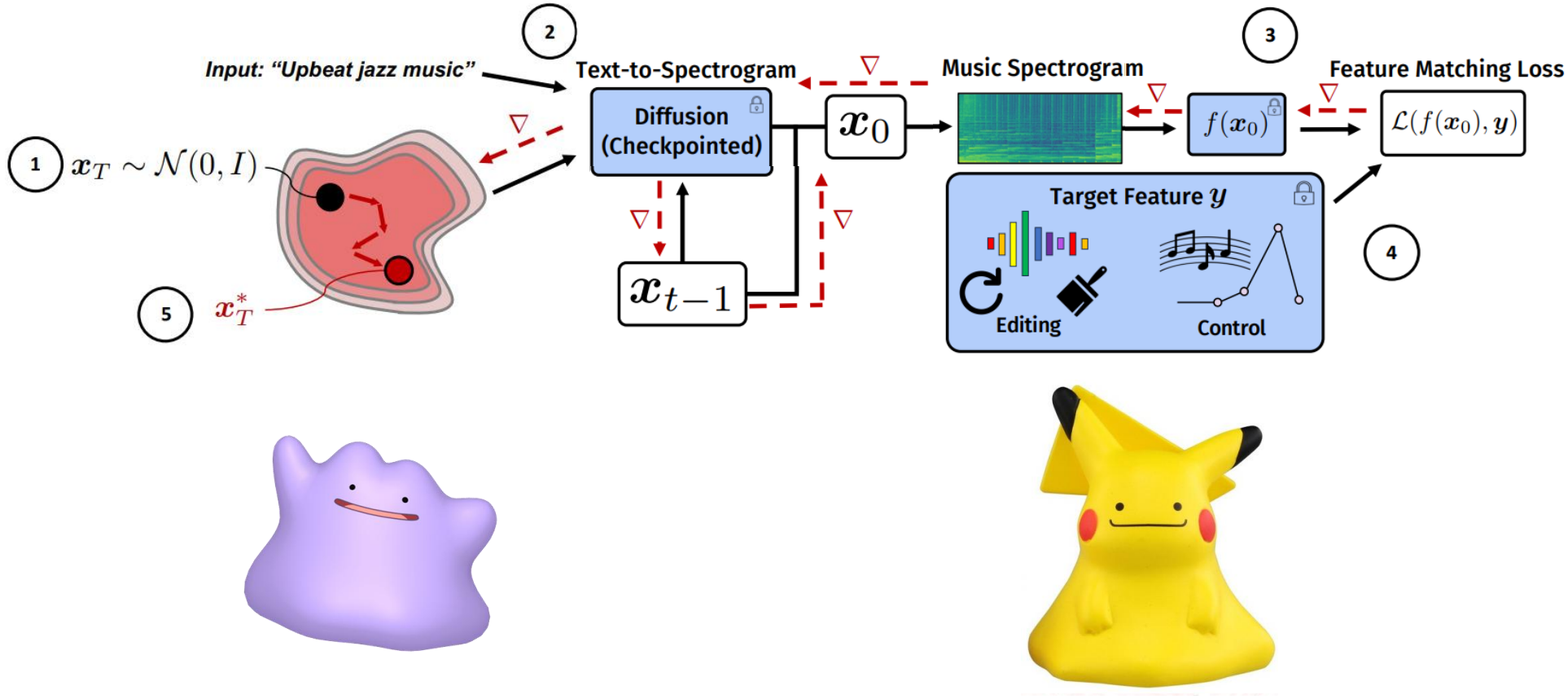
Inference-time Control: DITTO (Novack et al., 2024)



(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.

Inference-time Control: DITTO (Novack et al., 2024)

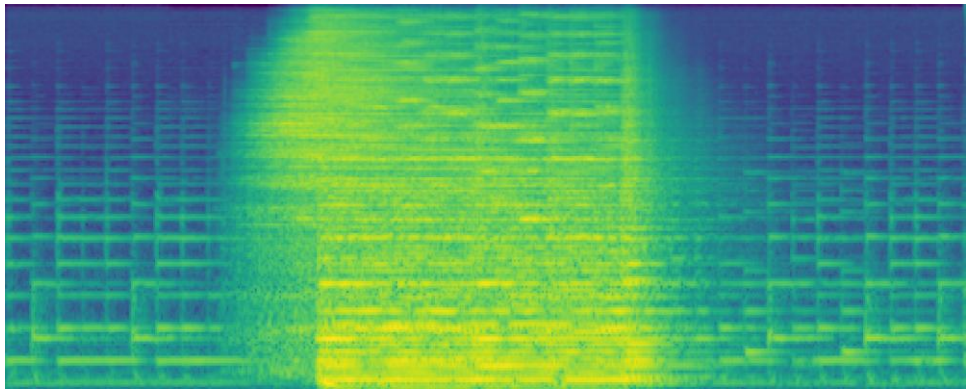


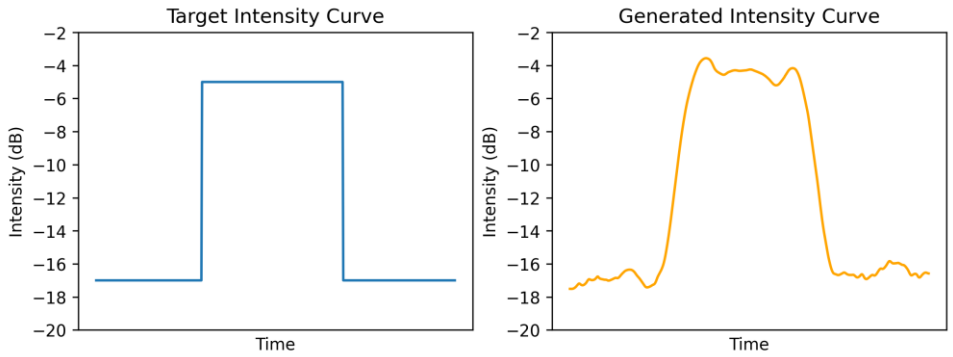
(Source: Novack et al., 2024)

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

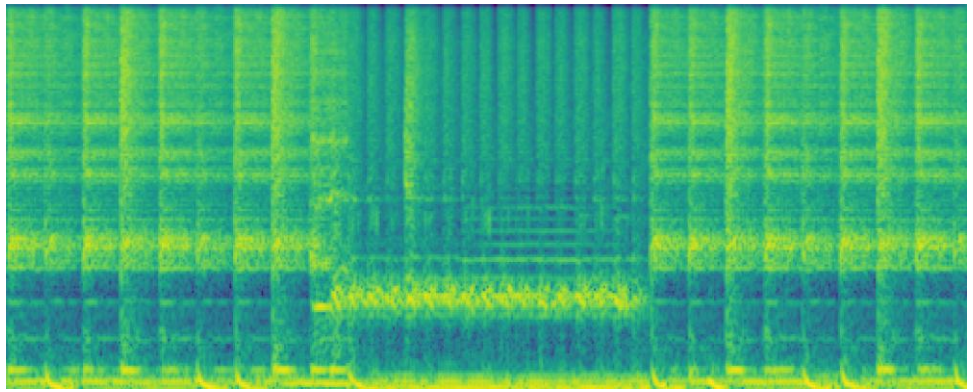
Inference-time Control: DITTO (Novack et al., 2024)

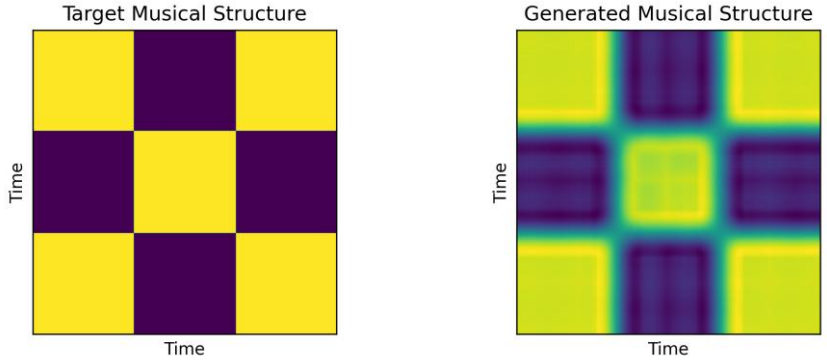
Intensity control





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.

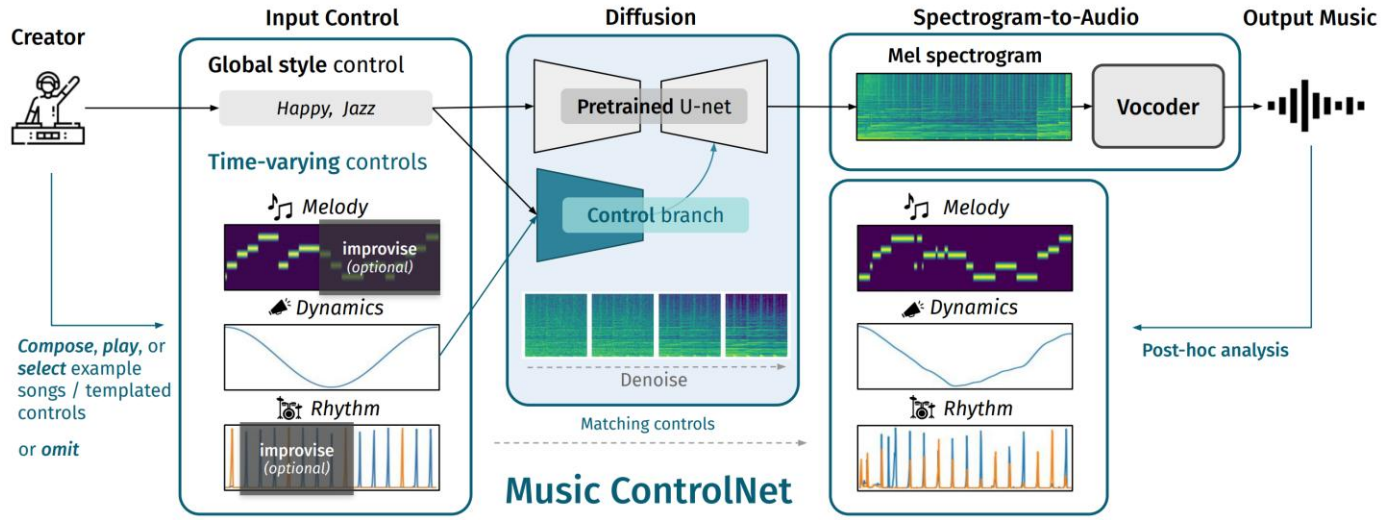
Inference-time Control: DITTO (Novack et al., 2024)

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

Music ControlNet vs DITTO

Music ControlNet

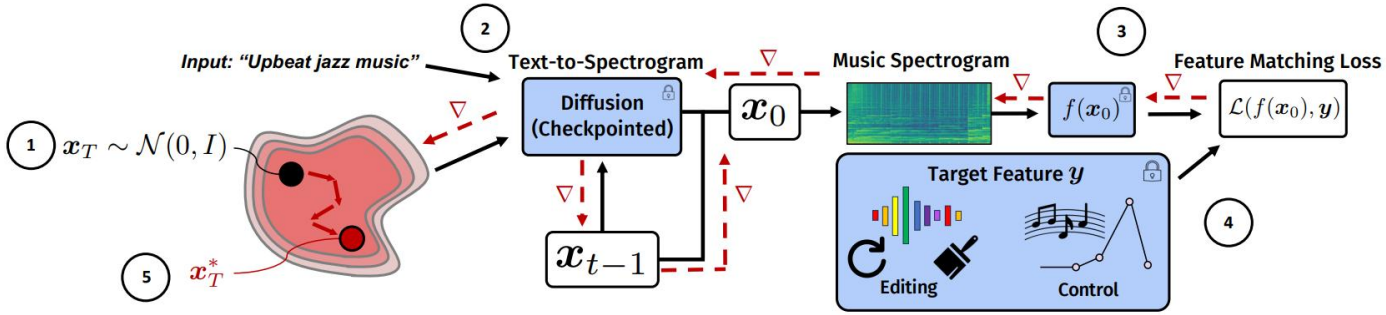
Needs some training!



(Source: Wu et al., 2024)

DITTO

No training needed!

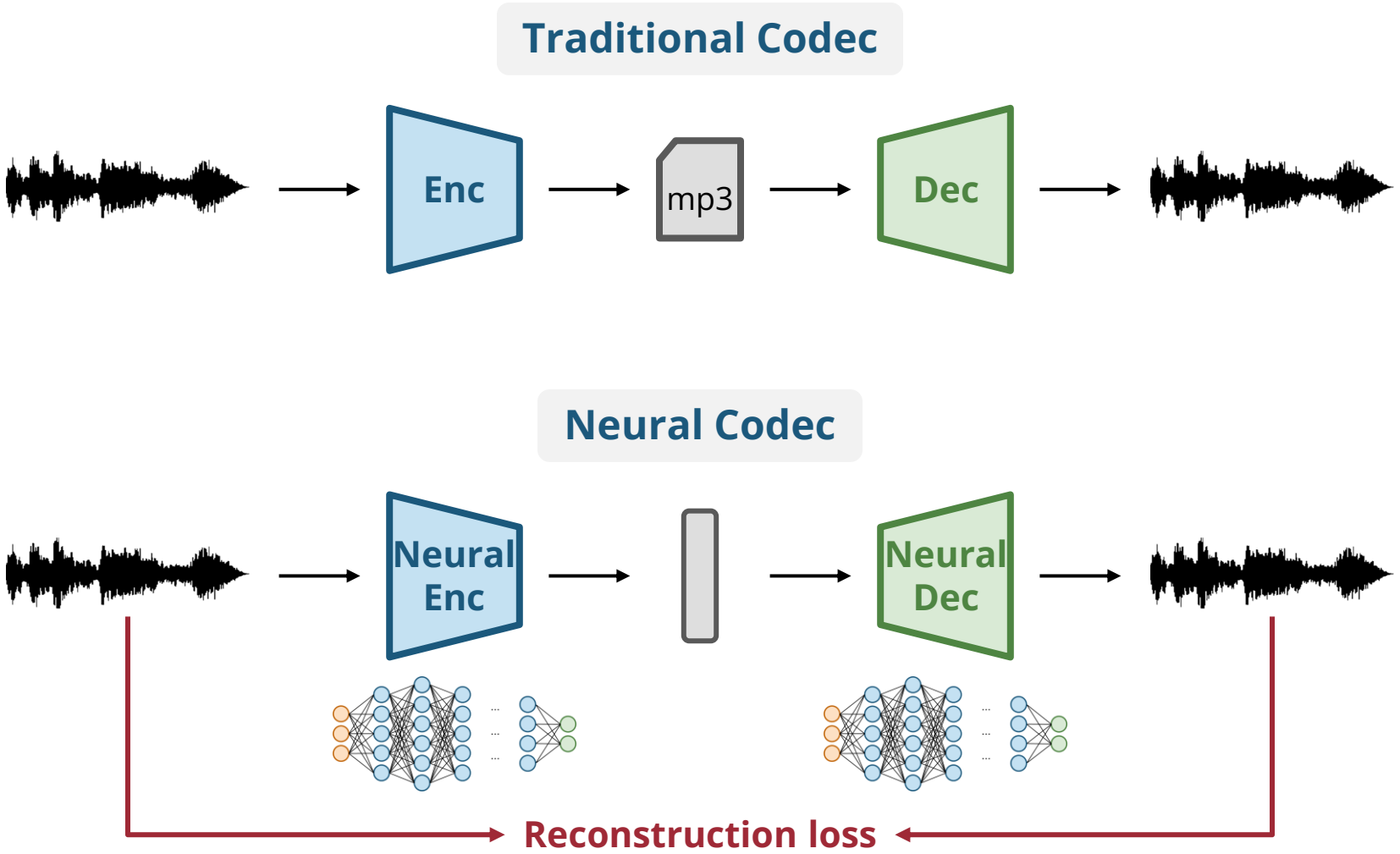


(Source: Novack et al., 2024)

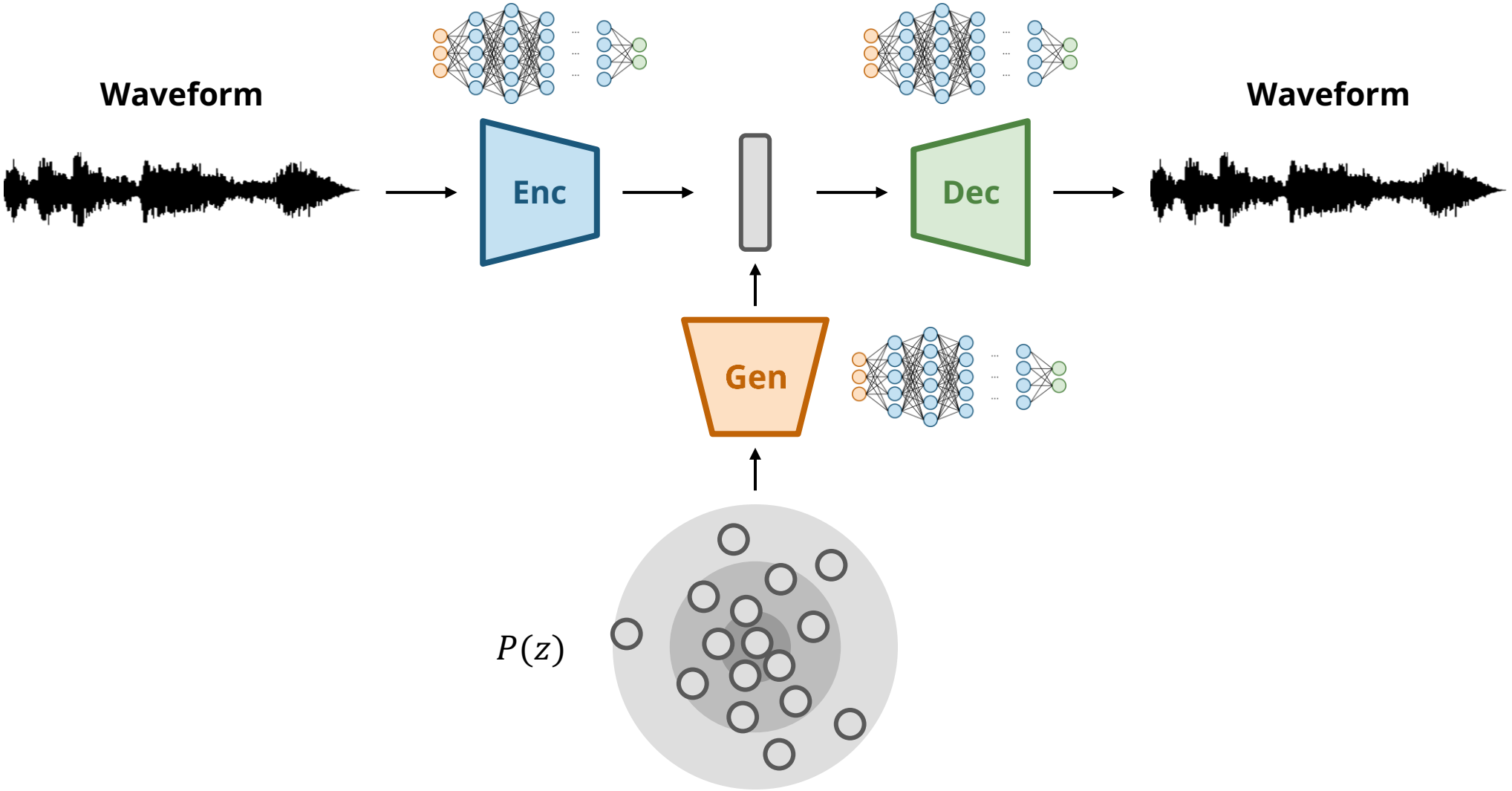
Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan, "Music ControlNet: Multiple Time-varying Controls for Music Generation," *TASLP*, 2024.
Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan, "DITTO: Diffusion Inference-Time T-Optimization for Music Generation,"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

Recap

Neural Codec

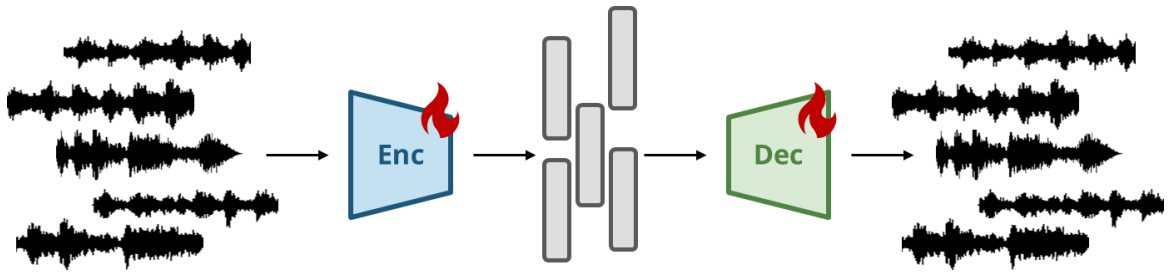


Latent-based Audio Synthesis

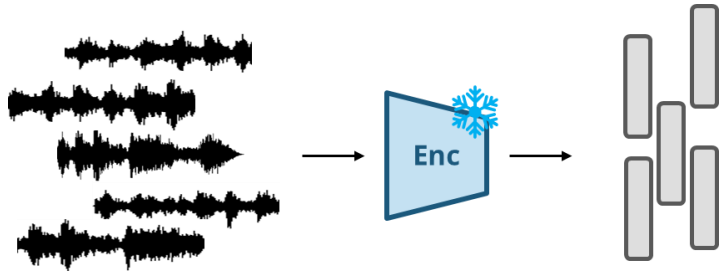


Pipeline

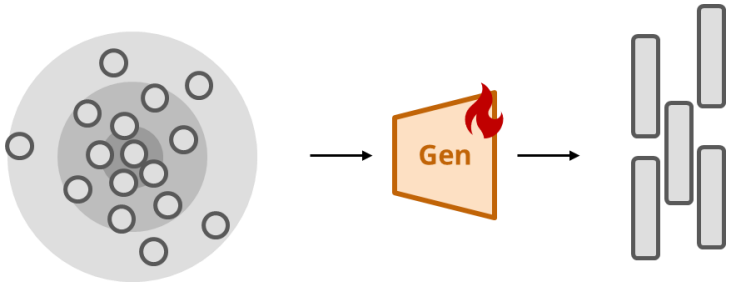
Step 1: Train an Autoencoder



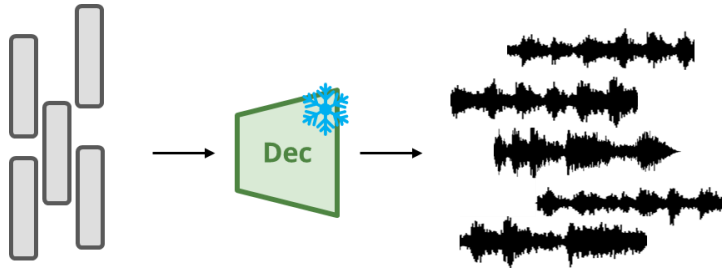
Step 2: Compute the Latent Vectors



Step 3: Train a Latent Generative Model

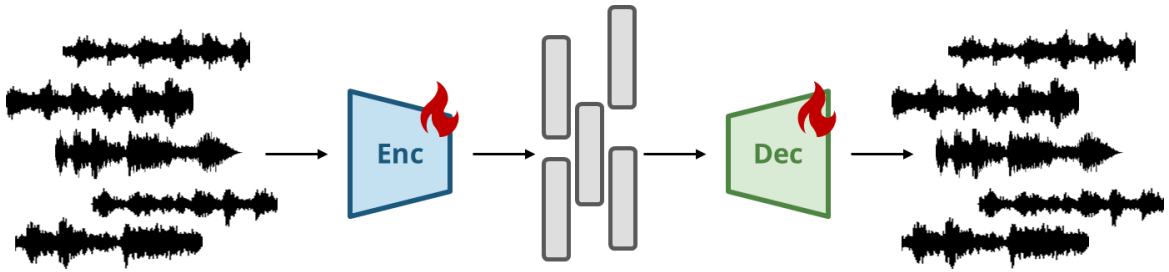


Step 4: Decode the Latent Vectors

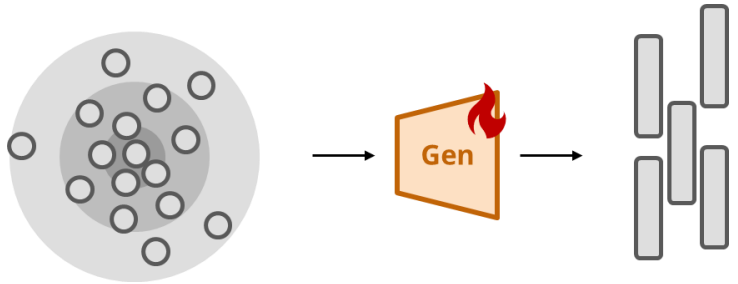


Training

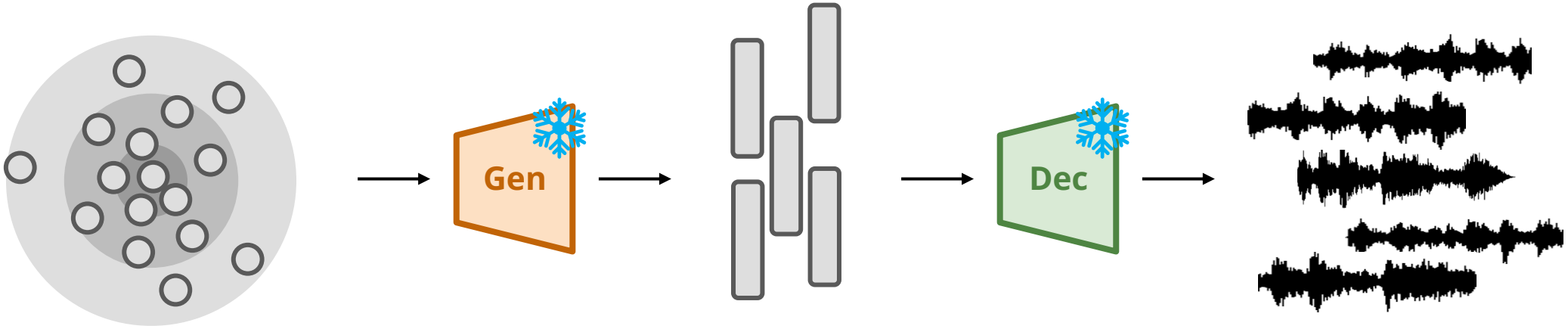
Autoencoder



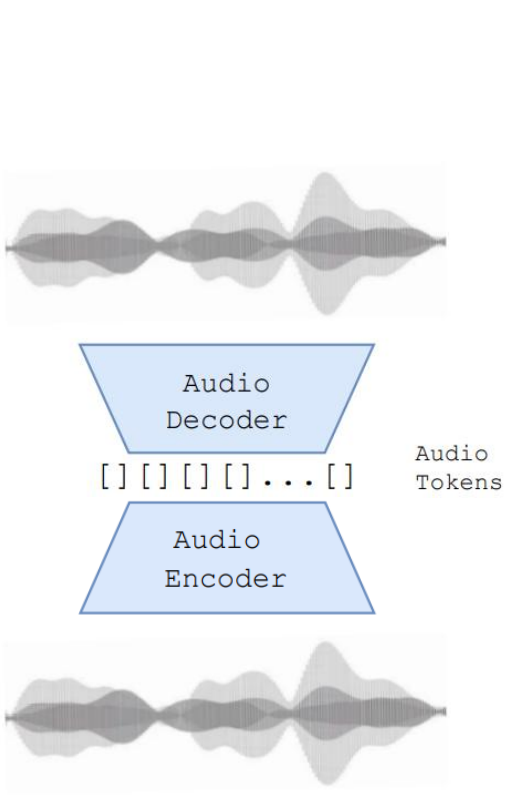
Latent Generative Model



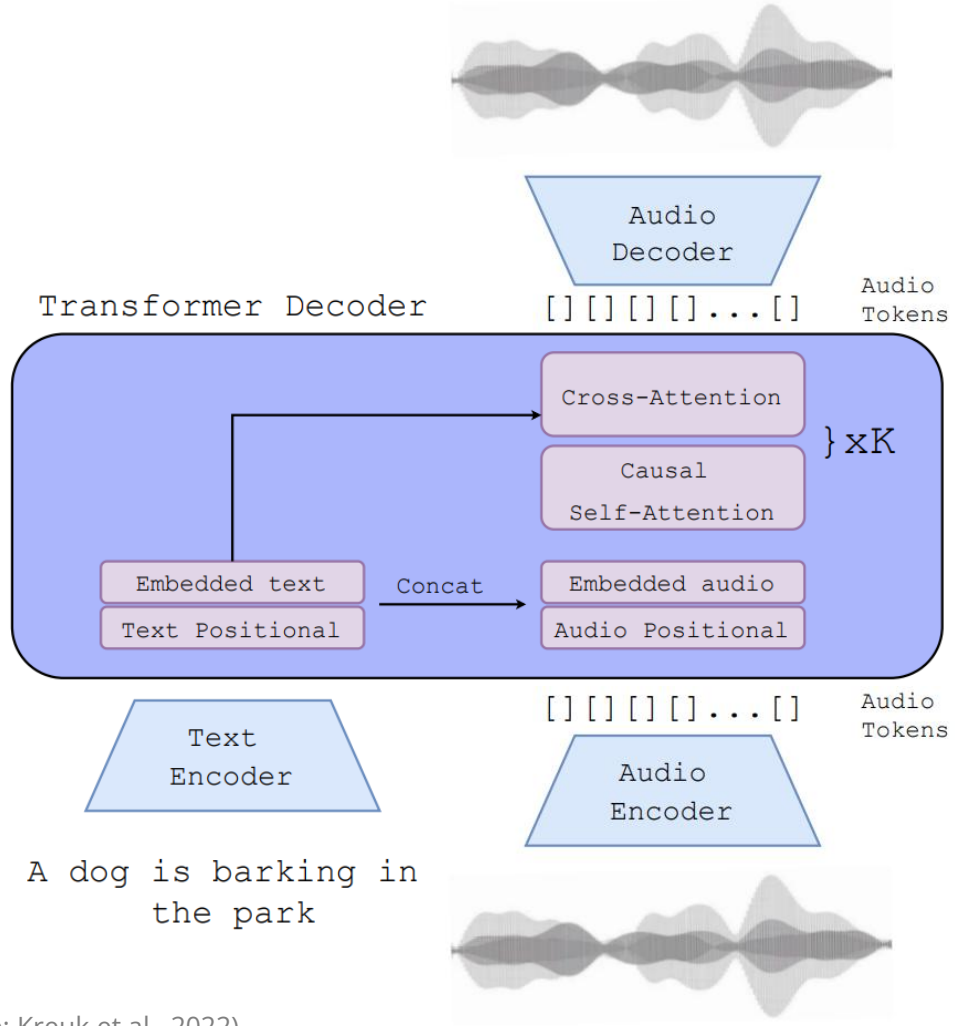
Inference



AudioGen (Kreuk et al., 2023)



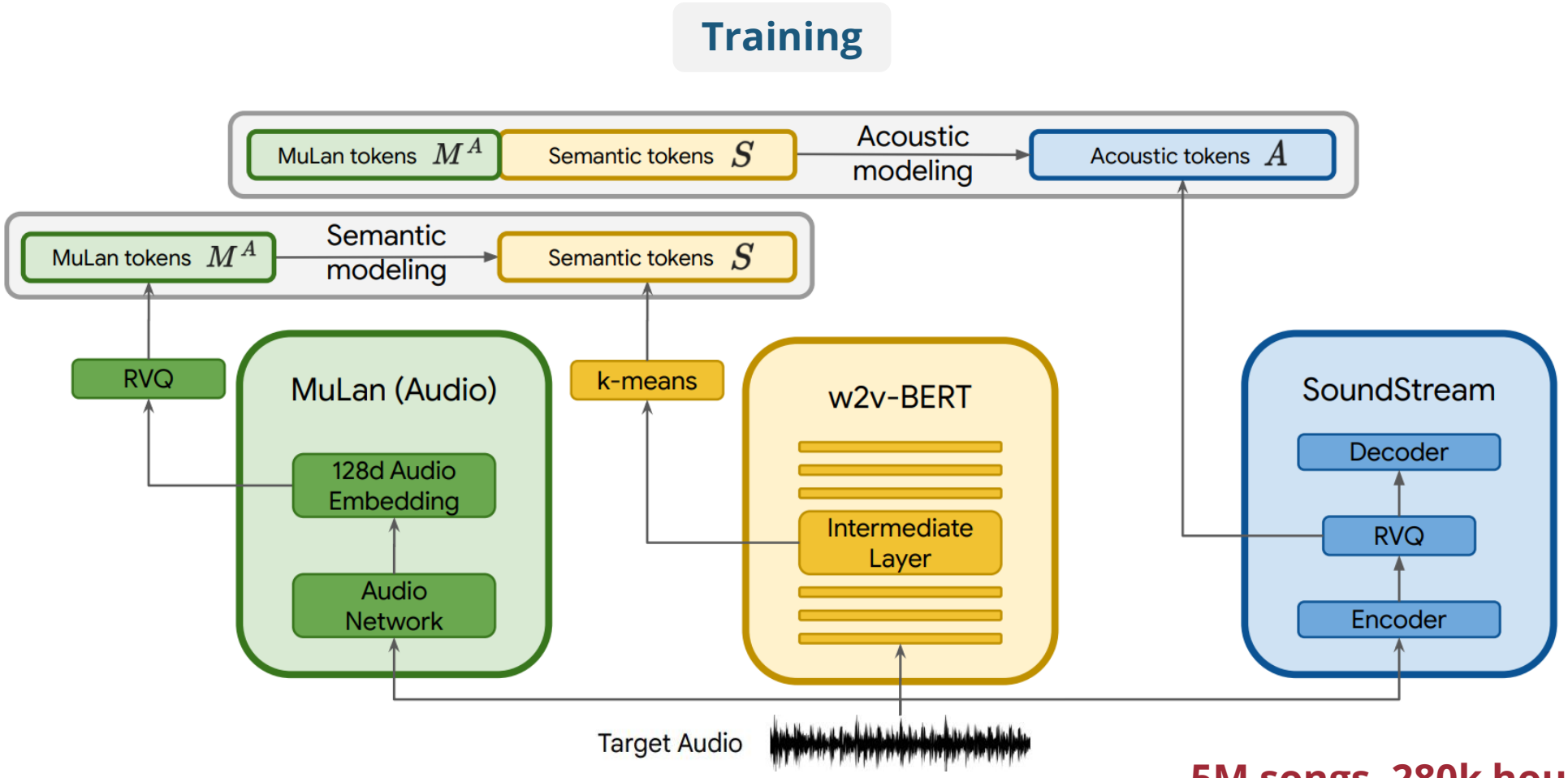
4k hours
(speech, music, sound effects)



(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.

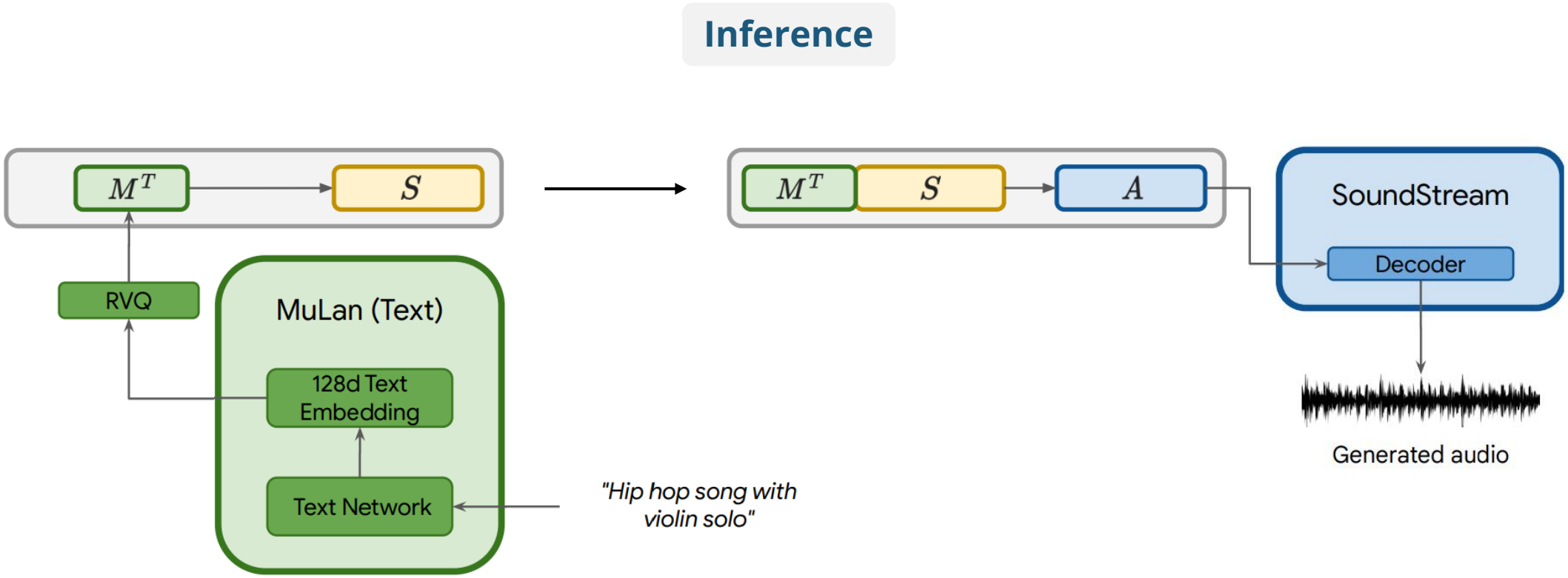
MusicLM (Agostinelli et al., 2023)



(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)

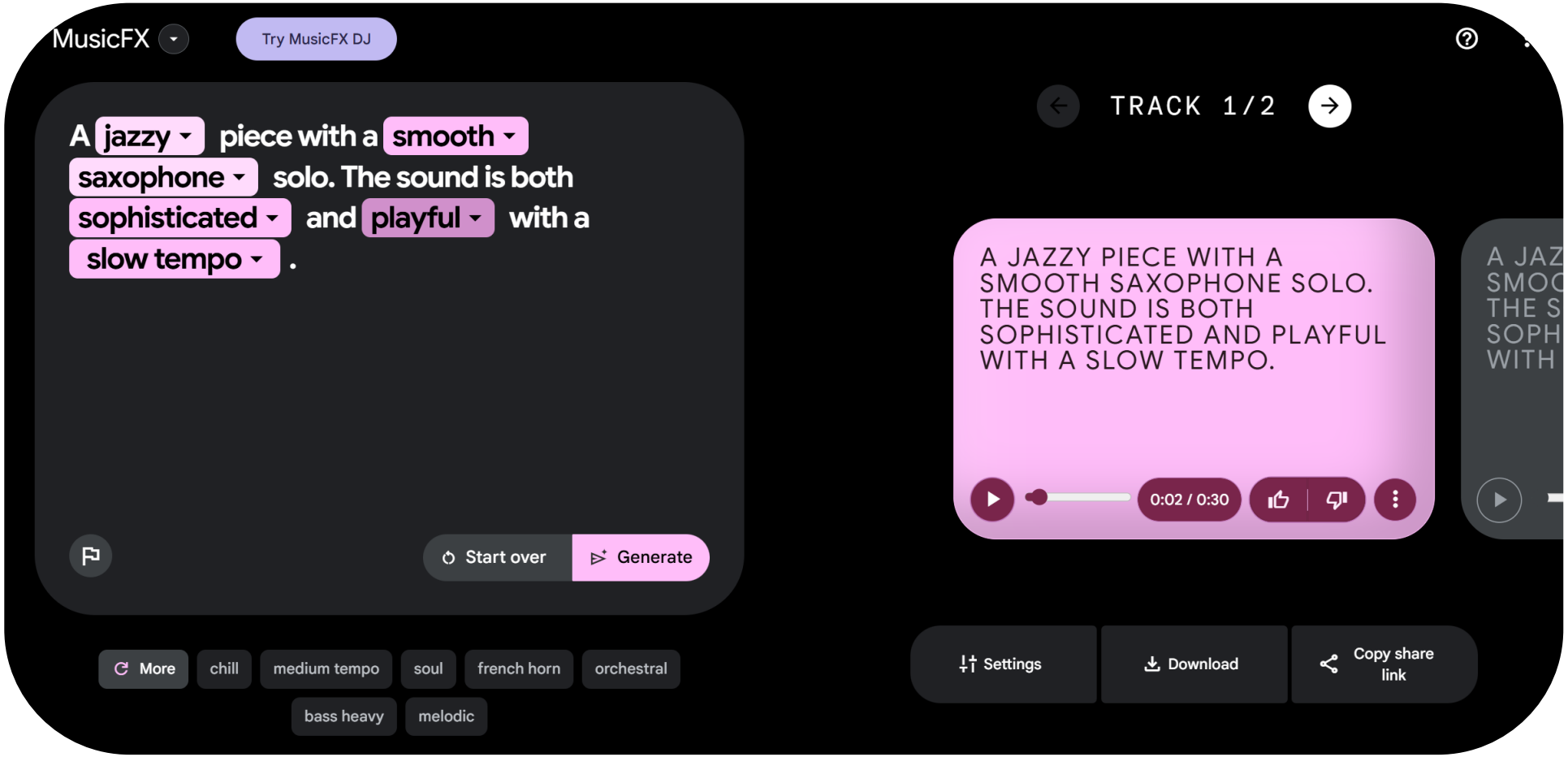


(Source: Agostinelli et al., 2022)

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

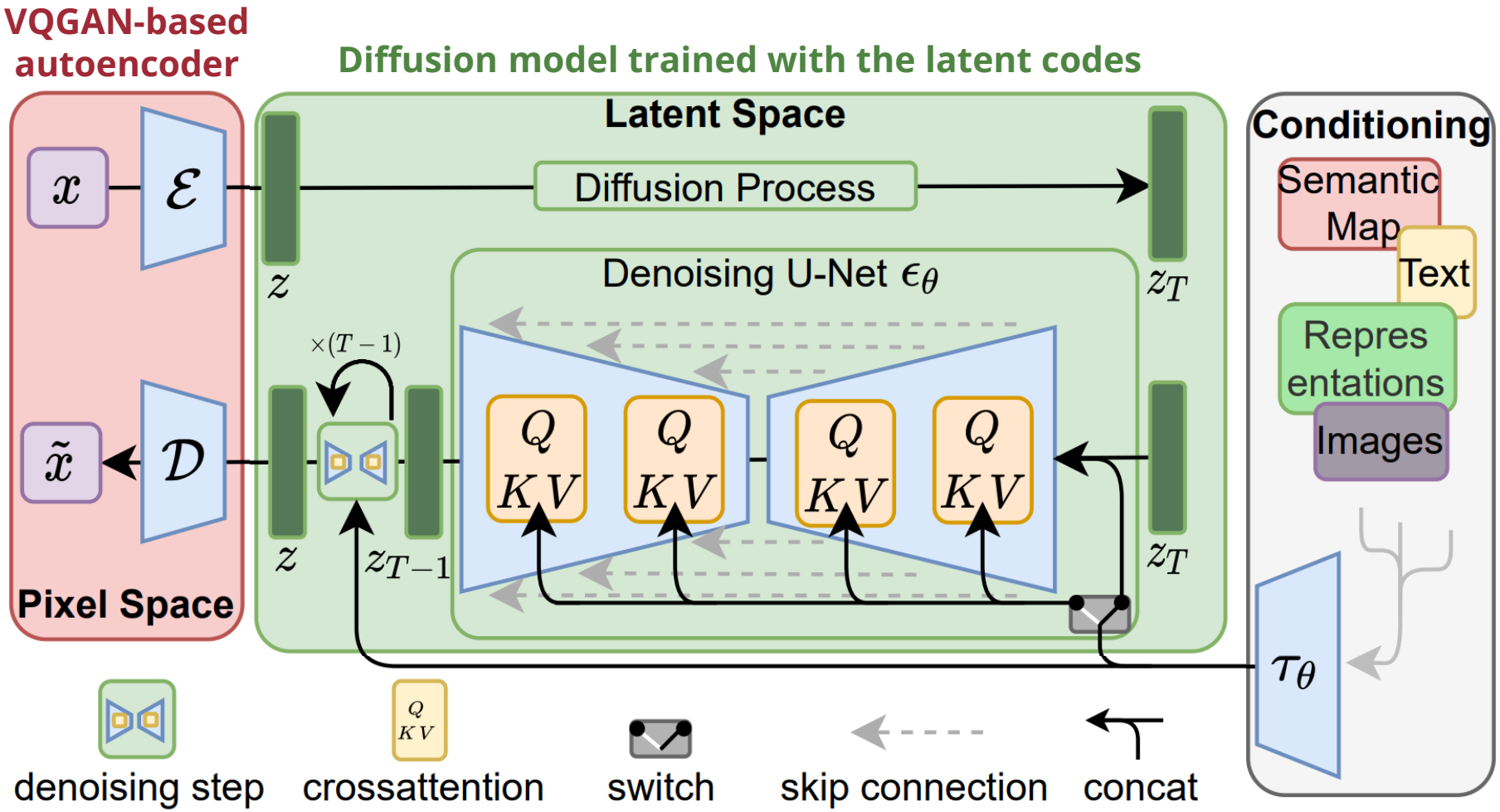
Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 2023.

Google's Music FX (2024)



aitestkitchen.withgoogle.com/tools/music-fx

Latent Diffusion Models (LDMs)

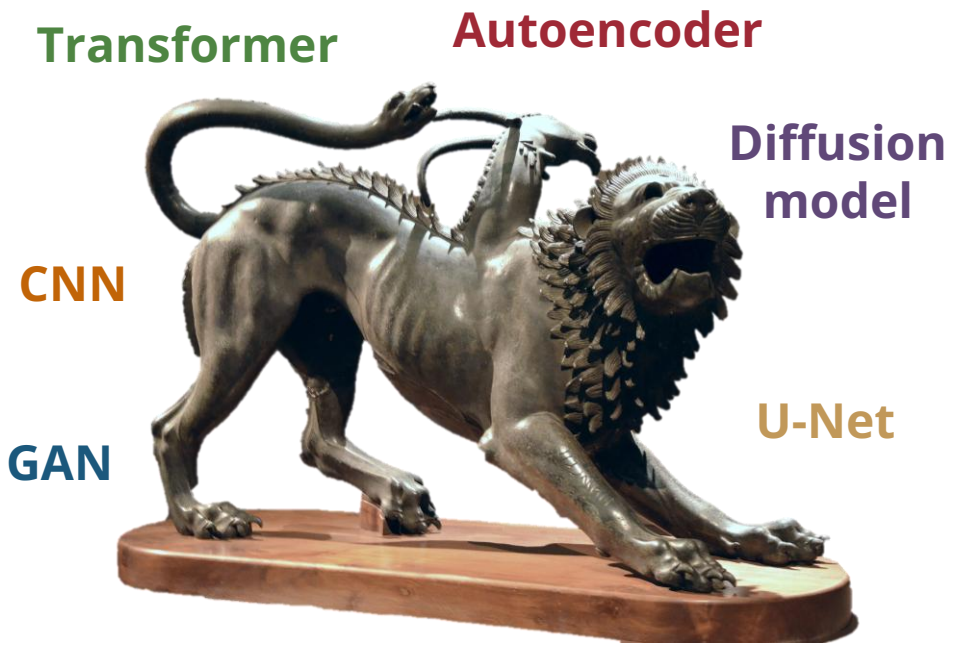


(Source: Rombach et al., 2022)

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

Latent Diffusion Model is a Chimera

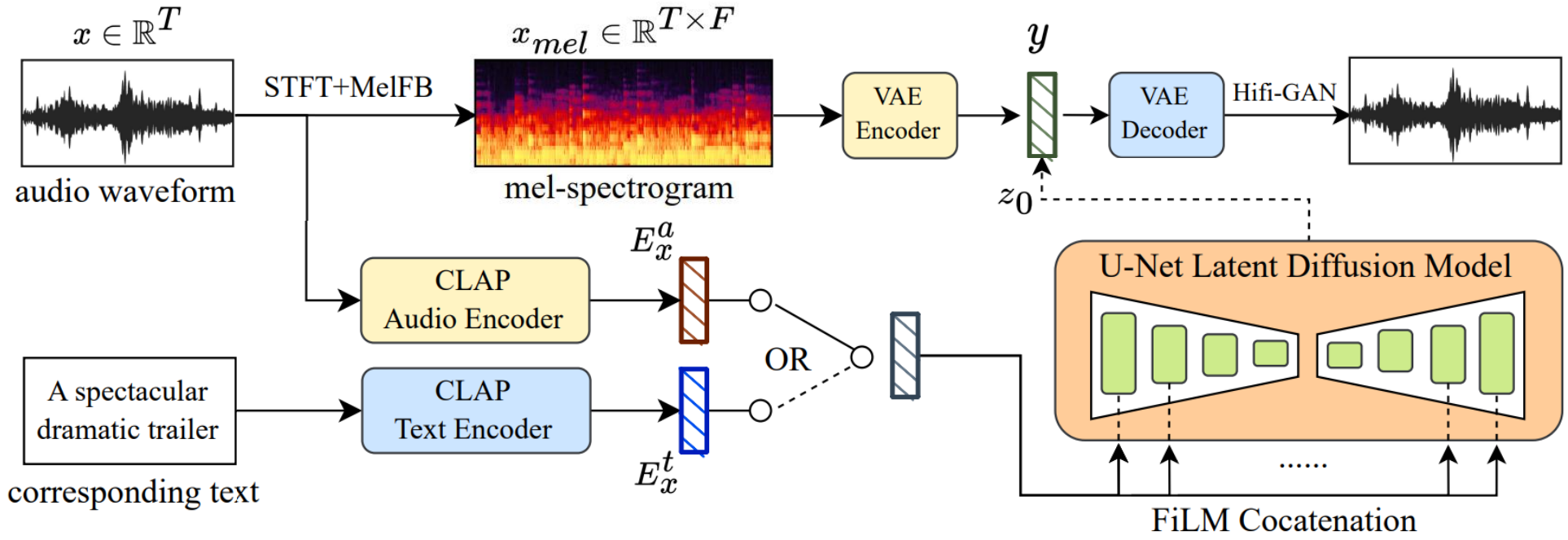
- **A neural codec**
  - An CNN-based autoencoder
  - Trained with a GAN-like adversarial loss
- **Diffusion model in the latent space**
  - A denoising U-Net
- **A conditioning module**
  - Transformer-like cross-attention mechanism



(Source: Raddato via worldhistory.org)

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

MusicLDM (Chen et al., 2023)

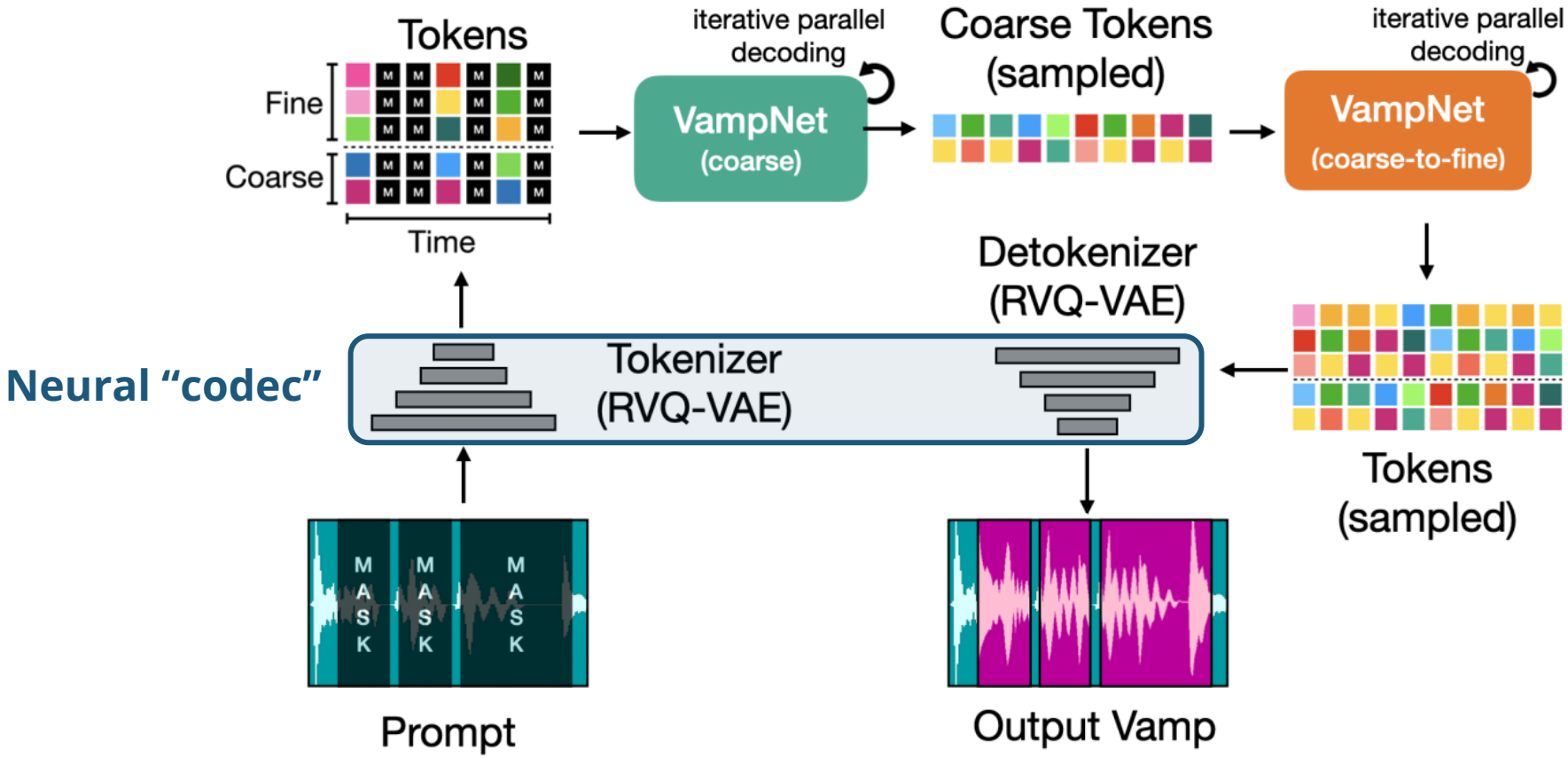


(Source: Ke et al., 2023)

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.

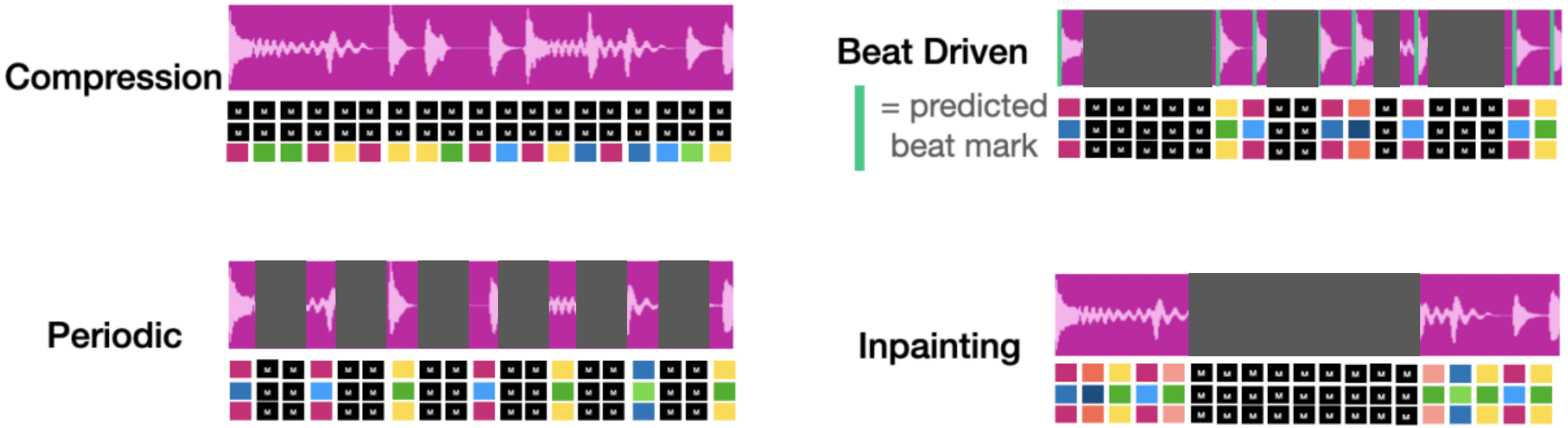
VampNet (Garcia et al., 2023)



(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)



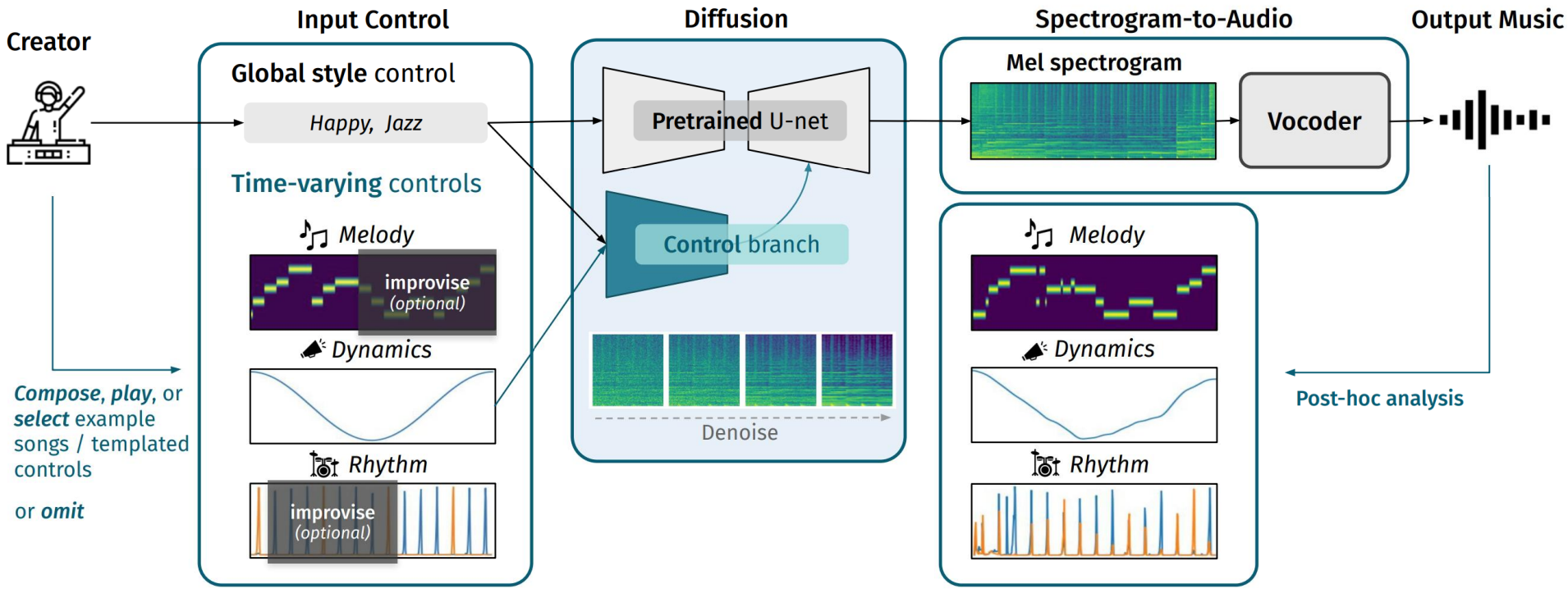
(Source: Garcia et al., 2023)

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

`unloop` (Garcia et al., 2023)

youtu.be/yzBI8Vcjd2s & github.com/hugofloresgarcia/unloop

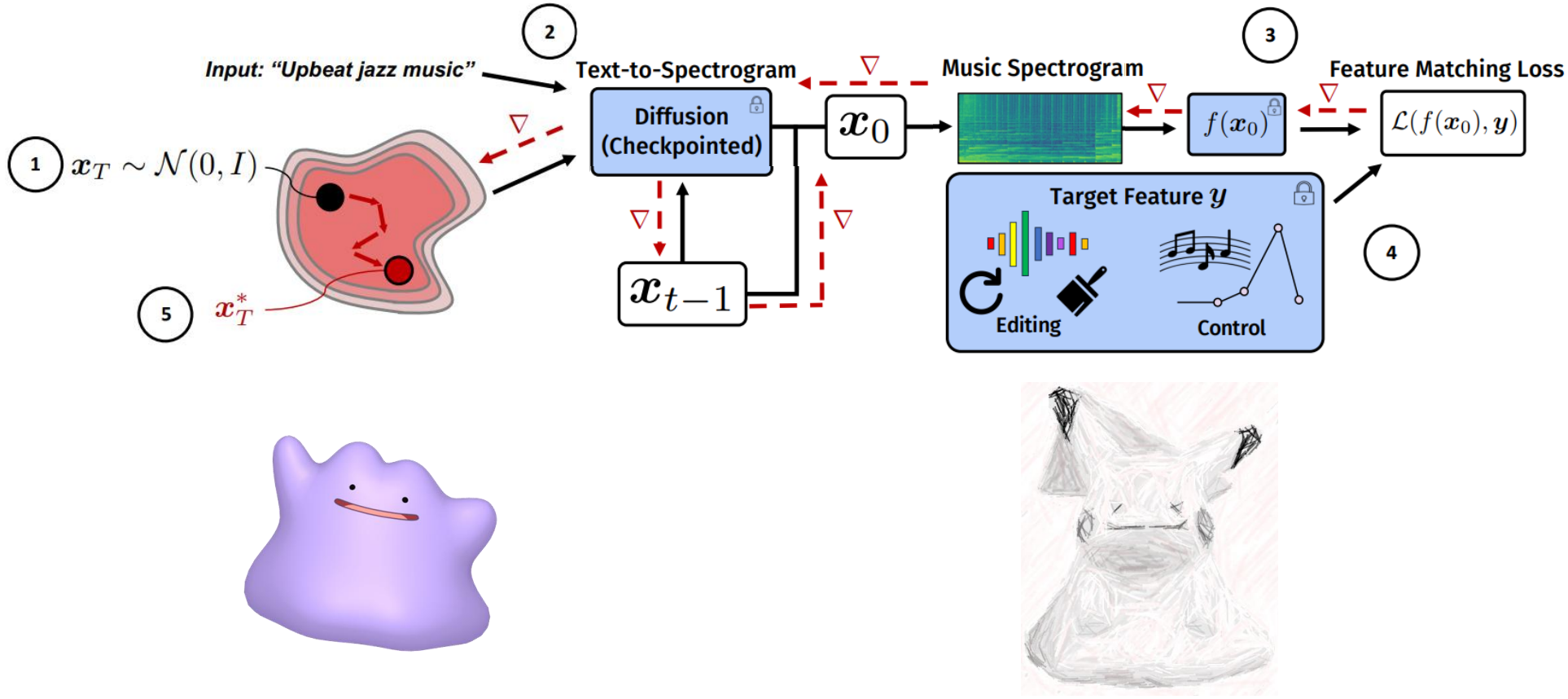
Music ControlNet (Wu et al., 2024)



(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.

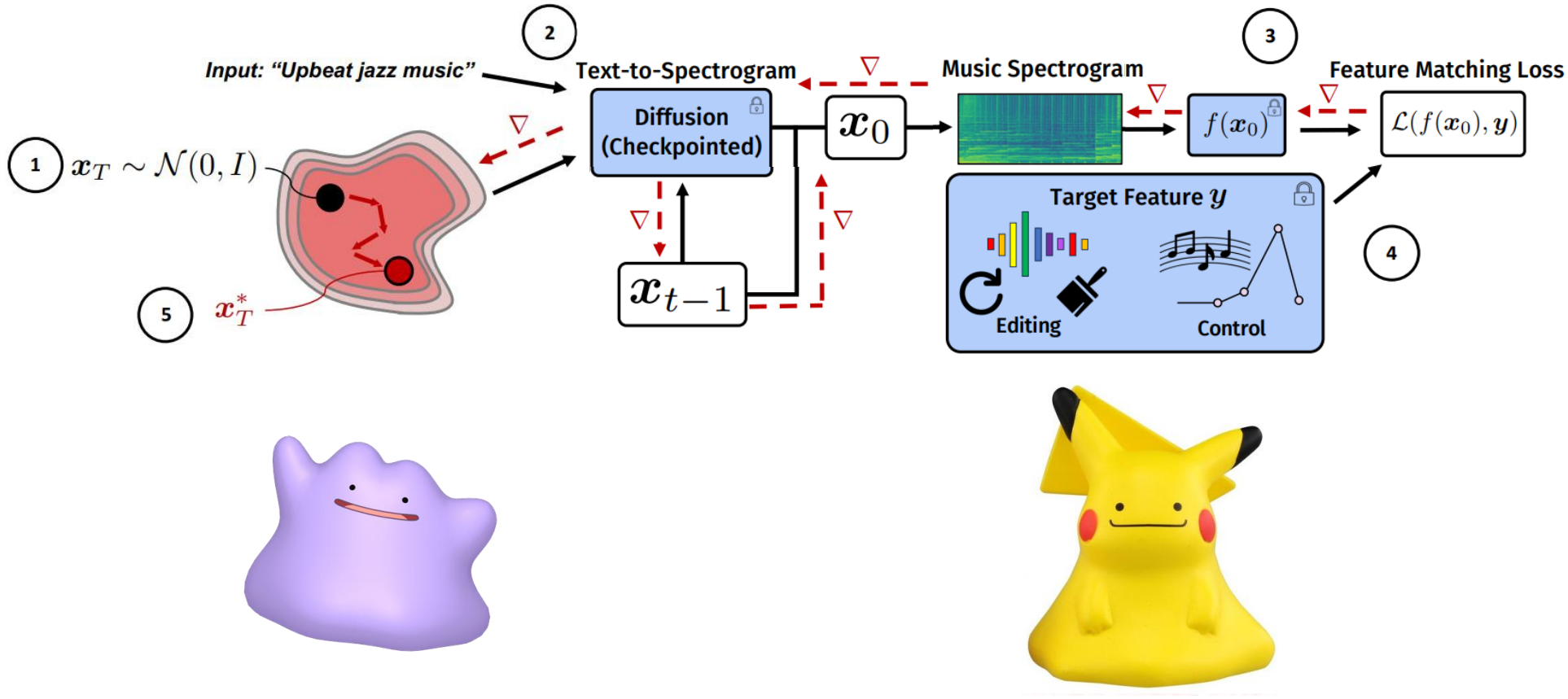
Inference-time Control: DITTO (Novack et al., 2024)



(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.

Inference-time Control: DITTO (Novack et al., 2024)



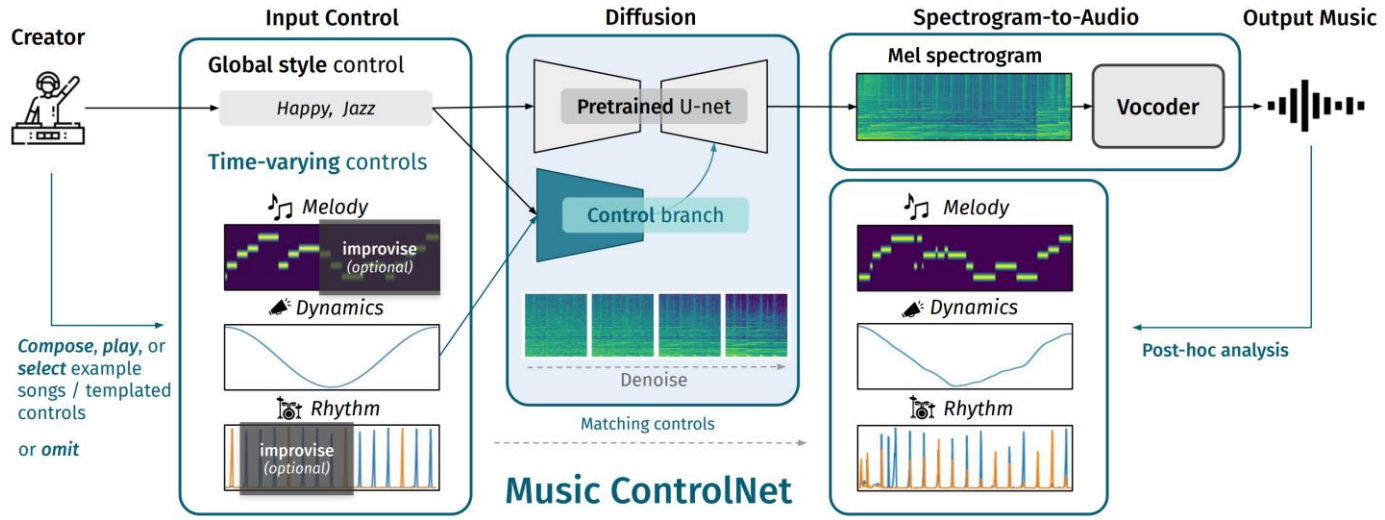
(Source: Novack et al., 2024)

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

Music ControlNet vs DITTO

Music ControlNet

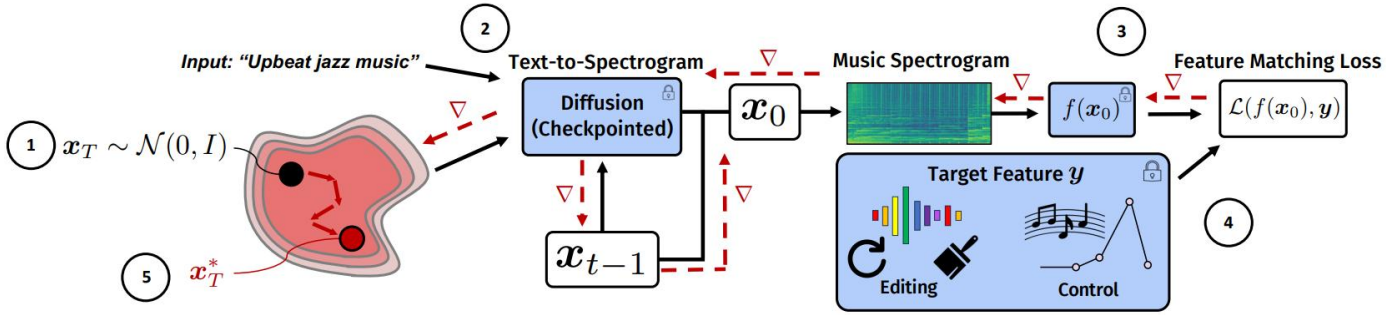
Needs some training!



(Source: Wu et al., 2024)

DITTO

No training needed!

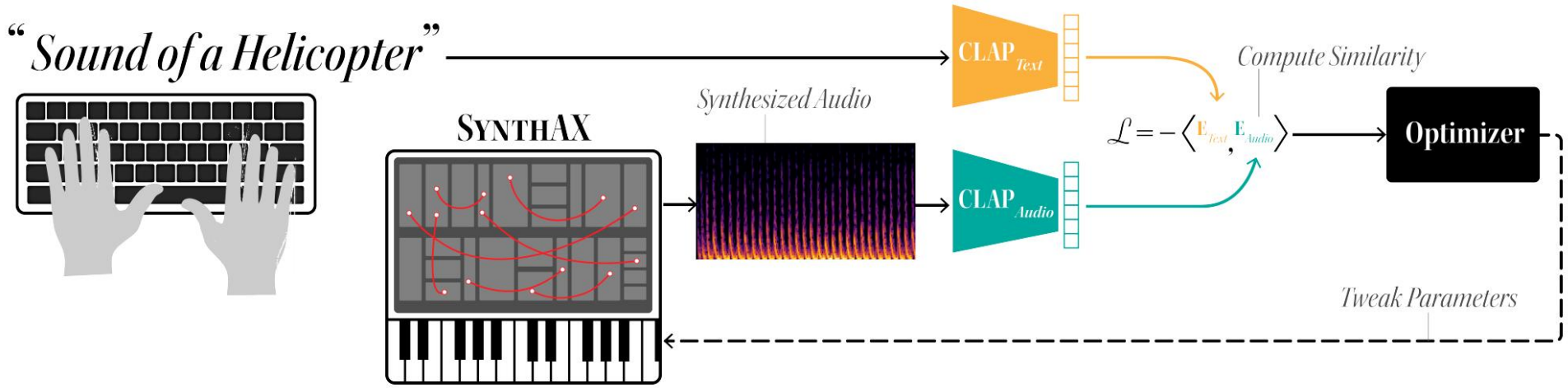


(Source: Novack et al., 2024)

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.

Next Lecture

Multimodal Systems & Music Production



(Source: Cherep et al., 2024)

ctag.media.mit.edu